Proceedings of the Fourteenth Dubrovnik MATH/CHEM/COMP International Course & Conference

The Fourteenth Dubrovnik International Course & Conference on the Interfaces among Mathematics, Chemistry and Computer Sciences was held in Dubrovnik, Croatia, on June 21-26, 1999, under the joint auspices of the Inter-University Centre, Dubrovnik, Croatia, Rugjer Bošković Institute, Zagreb, Croatia, Universities of Zagreb & Split, Croatia, International Society for Mathematical Chemistry, and International Society for Theoretical Chemical Physics.

FOREWORD

The MATH/CHEM/COMP (MCC) meetings have started at the Inter-University Centre in Dubrovnik in 1986 and soon have become a widely recognized forum to exchange the latest ideas on combinatorial, topological and computational aspects of chemistry and physics. The above topics have been presented through three out of six days of the MCC'99, namely at *Sessions on Theory and Applications of Mathematics in Chemistry, Computational and Experimental Chemistry*, and *Computations in Chemistry and Physics*. In order to keep the pace with the recent subjects of wider interest to chemical and pharmaceutical community the *Session of Drug and Vaccine Modelling* has been organized, as well as the *Session on Combinatorial Chemistry and Combinatorics in Chemistry*. Beside that, the third day of the meeting has accommodated the *Fifth Croatian Meeting on Fullerenes*.

These new trends, as well as traditional topics of the meeting are partially reflected in the present issue of the *Journal of Chemical Information and Computer Sciences*. Other papers from the meeting will appear as the special issues of journals *Croatica Chemica Acta* and *Fullerene Science and Technology*.

The MCC meetings encompass also courses usually held in the afternoon sessions, but the related materials are distributed only locally. However, the conference contributions are regularly published, which up to now have resulted in 15 special issues in various international journals. The papers presented at the MCC'97 have also appeared in this journal (and have been solicited by Ante Graovac, Dejan Plavšić and Dražen Vikić-Topić), but remained scattered through its issues of the year 1998. This time we have collected them in one issue, and special thanks for that go to Prof. Milne, editor of the *Journal of Chemical Information and Computer Sciences*.

Guest Editors: Ante Graovac (Zagreb & Split), Dejan Plavšić (Zagreb), and Dražen Vikić-Topić (Zagreb)

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1999	Conference Proceedings

4. TITLE AND SUBTITLE

MATH/CHEM/COMP'99

5. FUNDING NUMBERS

F61775-99-WF039

6. AUTHOR(S)

Conference Committee

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The R. Boskovic Institute
R-10 001
Zagreb POB 1016
Croatia

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

EOARD
PSC 802 BOX 14
FPO 09499-0200

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

CSP 99-5039

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

A

13. ABSTRACT (Maximum 200 words)

The Final Proceedings for MATH/CHEM/COMP'99, 21 June 1999 - 26 June 1999

This is an interdisciplinary conference. Topics include Discrete mathematics, especially combinatorics, graph and group theory, and computational methods as applied to chemistry, physics and material sciences. The meeting will contribute to better understanding of how the properties of various compounds and materials depend on their structure.

14. SUBJECT TERMS

EOARD, Computers, Mathematics, Chemistry

15. NUMBER OF PAGES

172

16. PRICE CODE

N/A

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

On the Relation between W'/W Index, Hyper-Wiener Index and Wiener Number

Dejan Plavšić [*,a], Nella Lerš [a], and Katica Sertić-Bionda [b]

[a]The Ruđer Bošković Institute, P.O.B. 1016, HR-10001 Zagreb,

The Republic of Croatia

[b]Faculty of Chemical Engineering and Technology, University of Zagreb,

HR-10000 Zagreb,The Republic of Croatia

It is shown analytically that the W'/W index, the hyper-Wiener index and the Wiener number are closely related graph-theoretical invariants for acyclic structures. A general analytical expression for the hyper-Wiener index of a tree is derived too.

1. INTRODUCTION

Properties of a molecule are a consequence of a complicated interplay of its topology (atomic connectivity), metric characteristics (bond lengths, valence and torsion angles) as also detailed dynamics of electrons and nuclei.[1] Within many classes of compounds, the variations of molecular metric and electronic structure are small. Hence, changes of many of molecular properties in these classes may be considered as only topology conditioned.[2-9] Molecular topology can be represented by a (molecular) graph being abstract, essentially non-numerical mathematical object.[10-15] In order to perform quantitative topology-property/activity studies of molecules it is necessary to quantify the structural

information contained in the corresponding graphs. The characterization of a graph is usually carried out by means of graph invariants[10] (topological indices[16]).

The representation of a molecule by a topological index entails a considerable loss of information concerning the molecular structure. Hence, chemists are permanently in pursuit of novel topological indices which would improve the graph-theoretical characterization of molecular structure and enable to make better and easily interpretable regression models of topology-property/activity relationships. In the framework of this effort Randić has recently put forward a novel bond additive molecular descriptor, the W'/W index.[17,18] The index is the sum of graphical bond orders[17,18] of all edges in a graph calculated by means of the Wiener number.[19] He also tested the W'/W index in the framework of the single variable linear regression model by examining the van der Waals areas of heptanes[17,20] and some 20 molecular properties of octane isomers.[21] The close resemblance in quality between the regressions with the W'/W index as predictor variable and regressions based on the hyper-Wiener index[20] as well as on the Wiener number indicates that these indices encode the very similar information on topology of acyclic structures. We have investigated the intercorrelation of these three indices on heptane, octane and nonane isomers and found in all cases high value of the coefficient of determination ($r^2 > 0.996$) and strong linear intercorrelation. The plot of the W'/W index versus the hyper-Wiener index R for nonanes is illustrated in Figure 1. Such a behavior of these three indices hints that a formal relationship might exist between them.

In this article we will discuss the relationship between the W'/W index, the hyper-Wiener index and the Wiener number for acyclic structures.

2. DEFINITIONS

Wiener Number. The Wiener number, $W=W(G)$, of a connected undirected graph G with N vertices is defined as[16]

$$W = \sum_{i=1}^{N-1} \sum_{j>i}^{N} (\mathbf{D})_{ij} \tag{1}$$

where $(\mathbf{D})_{ij}$ denotes the element in the i-th row and j-th column of the distance matrix $\mathbf{D}$ of the graph G. The summation goes over all the entries above the main diagonal of $\mathbf{D}$. If G is a connected undirected acyclic graph (tree), T, with N vertices the Wiener number can also be expressed as[19]

$$W = \sum_{e_{ij}} C_{e_{ij}} \tag{2}$$

where

$$C_{e_{ij}} = {}^{i}X_{e_{ij}} (N - {}^{i}X_{e_{ij}}) \; ; \; 1 \leq {}^{i}X_{e_{ij}} \leq N-1$$

$$= {}^{i}X_{e_{ij}} \; {}^{j}X_{e_{ij}} \tag{3}$$

e_{ij} denotes the edge connecting vertices i and j of T. The summation in eq 2 goes over all edges in T. ${}^{i}X_{e_{ij}}$ and ${}^{j}X_{e_{ij}}$ in eq 3 denote the number of vertices of T on the side of the vertex i and on the side of the vertex j of the edge e_{ij}, respectively.

Graphical Bond Order $W_{e_{ij}}/W$. The graphical bond order $W_{e_{ij}}/W$ of an edge e_{ij} of a connected undirected graph G is defined by[17,18]

$$W_{e_{ij}}/W = \frac{W(G-e_{ij})}{W(G)} \tag{4}$$

where W(G) is the Wiener number of G and $W(G-e_{ij})$ denotes the Wiener number of the spanning subgraph $G-e_{ij}$ obtained from G by deleting the edge e_{ij}. $G-e_{ij}$ is connected if and only if G contains at least one ring and the edge e_{ij} is one of the edges making up ring(s). A disconnected $G-e_{ij}$ has two components, say G_1 and G_2, and the Wiener index in this case is by definition given by the expression[17,18]

$$W(G-e) = W(G_1) + W(G_2) \tag{5}$$

W'/W Index. The W'/W index of a connected undirected graph G is defined as:[17,18]

$$W'/W = \sum_{e_{ij}} W_{e_{ij}}/W = \frac{1}{W(G)} \sum_{e_{ij}} W(G-e_{ij}) \tag{6}$$

where the summation goes over all edges of G.

Hyper-Wiener Index. The hyper-Wiener index R was recently introduced by Randić for an acyclic structure.[20] The R index, R=R(T), of a tree T is defined as

$$R(T) = \sum_{p_{rs}} {}^r X_{p_{rs}} {}^s X_{p_{rs}} \tag{7}$$

where p_{rs} represents the path connecting vertices r and s of T. ${}^r X_{p_{rs}}$ and ${}^s X_{p_{rs}}$ denote the number of vertices of T on each side of the path p_{rs}, including r and s, respectively. The summation runs over all paths in T. Note, if paths of length one (edges) are the only paths taken into consideration, than eq 7 is reduced to eq 2. The original definition was extended so as to be applicable for all connected graphs.[22] The hyper-Wiener index, R=R(G), of a connected graph G with N vertices is defined as

$$R = \frac{1}{2} \sum_{i=1}^{N-1} \sum_{j>i}^{N} \left[(\mathbf{D})_{ij} + (\mathbf{D})_{ij}^2 \right] \tag{8}$$

3. RELATIONSHIP BETWEEN W'/W INDEX, HYPER-WIENER INDEX AND WIENER NUMBER

Let T be a connected undirected acyclic graph with N vertices and let e_{ij} be an edge of T, see Figure 2. The graph T-e_{ij} obtained from T by deleting the edge e_{ij} has two components, T_1 and T_2 with ${}^iX_{e_{ij}}$ and ${}^jX_{e_{ij}}$ vertices, respectively. Clearly, ${}^iX_{e_{ij}} + {}^jX_{e_{ij}}$ is equal to N. The Wiener number of the spanning subgraph T-e_{ij} of T is smaller than the Wiener number of T due to the absence of the contribution of the edge e_{ij} being equal to ${}^iX_{e_{ij}} {}^jX_{e_{ij}}$, and decrease in contributions of the remaining edges. An edge of T-e_{ij}, say e_{ab} (see Figure 2.) makes a contribution to $W(T$-$e_{ij})$ equal to the difference between the contribution of the corresponding edge in T to $W(T)$, ${}^aX_{e_{ab}} {}^bX_{e_{ab}}$, and the product ${}^aX_{e_{ab}} {}^jX_{e_{ij}}$. Hence, the difference between $W(T)$ and $W(T$-$e_{ij})$ is given by

$$W(T) - W(T\text{-}e_{ij}) = {}^iX_{e_{ij}} {}^jX_{e_{ij}} + {}^jX_{e_{ij}} \sum_{e_{ab} \in E_1} {}^aX_{e_{ab}} + {}^iX_{e_{ij}} \sum_{e_{cd} \in E_2} {}^dX_{e_{cd}} \tag{9}$$

on the condition that

$$(\mathbf{D})_{ia} > (\mathbf{D})_{ib} \tag{10}$$

and

$$(\mathbf{D})_{jd} > (\mathbf{D})_{jc} \tag{11}$$

${}^aX_{e_{ab}}$ ($ {}^dX_{e_{cd}}$) is the number of vertices of T_1 (T_2) on the side of the vertex a (d) of the edge e_{ab} (e_{cd}). E_1 and E_2 denote the sets of edges of the components T_1 and T_2, respectively. The summation in the second (third) term on the right hand side of eq 9 runs over all edges of the component T_1 (T_2) of the graph

$T-e_{ij}$.

The graphical bond order $W_{e_{ij}}/W$ of the edge e_{ij} of T can be obtained by dividing eq 9 by $W(T)$:

$$W_{e_{ij}}/W = 1 - \frac{1}{W(T)}\left({}^iX_{e_{ij}}\,{}^jX_{e_{ij}} + {}^jX_{e_{ij}}\sum_{e_{ab}\in E_1}{}^aX_{e_{ab}} + {}^iX_{e_{ij}}\sum_{e_{cd}\in E_2}{}^dX_{e_{cd}}\right) \quad (12)$$

This quantity represents the "importance" of the edge e_{ij} in T.

The W'/W index of T is the sum of graphical bond orders of all N-1 edges of T:

$$W'/W = N-1 - \frac{1}{W(T)}\sum_{e_{ij}}\Big({}^iX_{e_{ij}}\,{}^jX_{e_{ij}} + {}^jX_{e_{ij}}\sum_{e_{ab}\in E_1}{}^aX_{e_{ab}} +$$

$$ {}^iX_{e_{ij}}\sum_{e_{cd}\in E_2}{}^dX_{e_{cd}}\Big) \quad (13)$$

By noting that

$$\sum_{e_{ij}}\left({}^iX_{e_{ij}}\,{}^jX_{e_{ij}} + {}^jX_{e_{ij}}\sum_{e_{ab}\in E_1}{}^aX_{e_{ab}} + {}^iX_{e_{ij}}\sum_{e_{cd}\in E_2}{}^dX_{e_{cd}}\right) = 2\,R(T) - W(T) \quad (14)$$

the relationship between the W'/W index, the hyper-Wiener index and the Wiener number of T immediately follows:

$$W'/W = N - \frac{2\,R(T)}{W(T)} \quad (15)$$

Since 2R is equal to the sum of the (unnormalized) second moment of distance D_2 and the Wiener number,[22] the W'/W index can also be expressed as:

$$W'/W = N-1 - \frac{D_2(T)}{W(T)} \quad (16)$$

4. DERIVATION OF GENERAL EXPLICIT FORMULA FOR
HYPER-WIENER INDEX OF TREE

A closer examination of eqs 6 and 15 reveals that it is possible by means of them to arrive at the general explicit formula for calculation of the hyper-Wiener index of T. To wit, combining eqs 6 and 15 one can write

$$R(T) = \frac{N\,W - W'}{2} \tag{17}$$

where W' is the sum of the Wiener numbers of all the spanning subgraphs T-e of T. It is well-known that the Wiener number can be calculated in a number of ways.[16,19,23-34] If one selects the route[23]

$$W = \sum_{n=1}^{n \leq N-1} n \; {}^{n}p \tag{18}$$

where ${}^{n}p$ denotes the number of paths of length n in T, then the general explicit formula for calculation of R(T) can be derived in a rather simple way.

The total number of paths of a given length n in the spanning subgraphs T-e of T, ${}^{n}P$, is given by the expression:

$$^{n}P = \sum_{T-e} {}^{n}p_{T-e} = (N - n - 1) \; {}^{n}p \tag{19}$$

where ${}^{n}p_{T-e}$ is the number of paths of length n in T-e, and the summation runs over all the spanning subgraphs T-e of T.

W' of T is given by

$$W' = \sum_{n=1}^{n \leq N-1} n \; {}^{n}P = \sum_{n=1}^{n \leq N-1} n \, (N - n - 1) \; {}^{n}p \tag{20}$$

Combining eqs 17, 18 and 20 one obtains:

$$R(T) = \frac{1}{2} \sum_{n=1}^{n \leq N-1} (n^2 + n) \; {}^n p \tag{21}$$

To calculate the R index of T it is necessary and sufficient to have knowledge of ${}^n p$'s of T. The number of paths of a given length n in T can be calculated either by means of the recursive relation or using the general analytical expression for ${}^n p$ in T.[35] The recursive relation reads as:

$${}^n p = \sum_{(\mathbf{D})_{ij}=n-2} v_i v_j - 2 \, ({}^{n-1} p) - {}^{n-2} p \quad ; \quad n \geq 3 \tag{22}$$

where v_i is the valence of the vertex i, and $(\mathbf{D})_{ij}=n-2$ denotes that the topological distance between the vertices i and j is equal to n-2. The summation runs over all the pairs of vertices of T separated by the paths of length n-2. The initial conditions are

$${}^1 p = N - 1 \tag{23}$$

$${}^2 p = \frac{1}{2} \sum_i v_i^2 - N + 1 \tag{24}$$

The general analytical expression for ${}^n p$ in T reads as follows:

$${}^n p = \sum_{m=1}^{n-2+k} (-1)^{m+s} (n - m - 1 + t) \sum_{\substack{(\mathbf{D})_{ij}=m \\ i<j}} v_i v_j +$$

$$(-1)^n \left(\frac{n-1}{2} \sum_{(\mathbf{D})_{ij}=0} v_i v_j - N + 1 \right) \tag{25}$$

Note that $\sum_{(\mathbf{D})_{ij}=0} v_i v_j = \sum_i v_i^2$. The parameters k, s and t take the following values:

$$k = \begin{cases} 2 & \text{for } n = 1 \\ 1 & \text{for } n = 2 \\ 0 & \text{for } n \geq 3 \end{cases} \tag{26}$$

$$s = \begin{cases} 1 & \text{for } n = \text{odd} \\ 0 & \text{for } n = \text{even} \end{cases} \tag{27}$$

and

$$t = \begin{cases} 1 & \text{for } n = 1 \\ 0 & \text{for } n \geq 2 \end{cases} \tag{28}$$

Combining eqs 21 and 25 one obtains the general explicit formula for calculation of the R index of T

$$R(T) = \frac{1}{2} \sum_{n=1}^{n \leq N-1} (n^2 + n) \left\{ \sum_{m=1}^{n-2+k} (-1)^{m+s} (n - m - 1 + t) \sum_{\substack{(D)_{ij}=m \\ i<j}} v_i v_j + (-1)^n \left(\frac{n-1}{2} \sum_{(D)_{ij}=0} v_i v_j - N + 1 \right) \right\} \tag{29}$$

where the parameters k, s and t take the values from eqs 26, 27 and 28, respectively. The application of the formula is illustrated for the hydrogen suppressed graph of 2,2,3,4-tetramethylpentane in Figure 3.

A special case of an acyclic graph is the path graph, P_N. Eq 29 in case of P_N takes a rather simple form

$$R(P_N) = \frac{1}{2} \sum_{n=1}^{n=N-1} (n^2 + n) (N - n) \tag{30}$$

whose summation gives the formula already derived by Lukovits.[36]

ACKNOWLEDGMENT

REFERENCES AND NOTES

(1) McWeeny, R. *Methods of Molecular Quantum Mechanics;* 2nd ed.; Academic Press: London, 1989.

(2) Rouvray, D. H. Predicting Chemistry from Topology. *Sci. Am.* **1986**, 254(3), 36-43.

(3) Turro, N. J. Geometrical and Topological Thinking in Organic Chemistry. *Angew. Chem. Int. Ed. Engl.* **1986**, 25, 882-901.

(4) Rouvray, D. H. The Modeling of Chemical Phenomena Using Topological Indices. *J. Comput. Chem.* **1987**, 8, 470-480.

(5) Randić, M. The Nature of Chemical Structure. *J. Math. Chem.* **1990**, 4, 157-189.

(6) Randić, M. Chemical Structure - What is She? *J. Chem. Educ.* **1992**, 69, 713-718.

(7) Rouvray, D. H. A Rationale for the Topological Approach to Chemistry. *J. Mol. Struct. (Theochem)* **1995**, 336, 101-114.

(8) Randić, M. On Characterization of Chemical Structure. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 672-687.

(9) Milne, G. W. Mathematics as a Basis for Chemistry. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 639-644.

(10) Harary, F. *Graph Theory;* Addison-Wesley: Reading, Massachusetts, 1972.

(11) Kier, L. B.; Hall, L. M. *Molecular Connectivity in Structure-Activity Analysis;* Research Studies Press Ltd.: Letchworth, England, 1986.

(12) Gutman, I.; Polansky, O. *Mathematical Concepts in Organic Chemistry;* Springer-Verlag: Berlin, 1986.

(13) Bonchev, D.; Rouvray, D. H. (Eds.) *Chemical Graph Theory. Introduction and Fundamentals;* Abacus Press, Gordon and Breach: New York, 1991.

(14) Trinajstić, N. *Chemical Graph Theory;* 2nd revised ed.; CRC Press: Boca Raton, FL, 1992.

(15) Balaban, A. T. (Ed.) *From Chemical Topology to Three-Dimensional Geometry:* Plenum Press: New York, 1997.

(16) Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, 44 , 2332-2339.

(17) Randić, M. Search for Optimal Molecular Descriptors. *Croat. Chem. Acta* **1991**, 64, 43-54.

(18) Randić, M.; Mihalić, Z.; Nikolić, S.; Trinajstić, N. Graphical Bond Orders: Novel Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, 34 , 403-409.

(19) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, 69 , 17-20.

(20) Randić, M. Novel Molecular Descriptor for Structure-Property Studies. *Chem. Phys. Lett.* **1993**, 211, 478-483.

(21) Randić, M. Comparative Regression Analysis. Regressions Based on a Single Descriptor. *Croat. Chem. Acta* **1993**, 66, 289-312.

(22) Klein, D. J.; Lukovits, I.; Gutman, I. On the Definition of the Hyper-Wiener Index for Cycle-Containing Structures. *J. Chem. Inf. Comput. Sci.* **1995**, 35 , 672-687.

(23) Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix and Molecular Branching. *J. Chem. Phys.* **1977**, 67, 4517-4533.

(24) Bersohn, M. A Fast Algorithm for Calculation of the Distance Matrix. *J. Comput. Chem.* **1983**, 4, 110-113.

(25) Canfield, E. R.; Robinson, R. W.; Rouvray, D.H. Determination of the Wiener Molecular Branching Index for the General Tree. *J. Comput. Chem.* **1985**, 6, 598-609.

(26) Barysz, M.; Plavšić, D.; Trinajstić, N. A Note on Topological Indices. *Math. Chem. (Mulcheim / Ruhr)* **1986**, 19, 89-116.

(27) Muller, W. R.; Szymanski, K.; Knop, J. V.; Trinajstić, N. An Algorithm for Construction of the Molecular Distance Matrix. *J. Comput. Chem.* **1987**, 8, 170-173.

(28) Mohar, B.; Pisanski, T. How to Compute the Wiener Index of a Graph. *J. Math. Chem.* **1988**, 2, 267-277.

(29) Senn, P. The Computation of the Distance Matrix and the Wiener Index for Graphs of Arbitrary Complexity with Weighted Vertices and Edges. *Comput. Chem.* **1988**, 12, 219-227.

(30) Lukovits, I. General Formulas for the Wiener Index. *J. Chem. Inf. Comput. Sci.* **1991**, 31, 503-507.

(31) Gutman, I.; Yeh, Y. N.; Lee, S. L.; Luo, Y. L. Some Recent Results in the Theory of the Wiener Number. *Indian J. Chem.* **1993**, 32A, 651-661.

(32) Plavšić, D.; Nikolić, S.; Trinajstić, N.; Klein, D. J. Relation between the Wiener Index and the Schultz Index for Several Classes of Chemical Graphs. *Croat. Chem. Acta* **1993**, 66, 345-353.

(33) Gutman, I. A New Method for the Calculation of the Wiener Number of Acyclic Molecules. *J. Mol. Struct. (Theochem)* **1993**, 285, 137-142.

(34) Nikolić, S.; Trinajstić, N.; Mihalić, Z. The Wiener Index: Development and Applications. *Croat. Chem. Acta* **1995**, 68, 105-129.

(35) Plavšić, D.; Šoškić, M.; Landeka, I.; Trinajstić, N. On the Relation between the P'/P Index and the Wiener Number. *J. Chem. Inf. Comput. Sci.* **1996**, 36 , 1123-1126.

(36) Lukovits, I. Formulas for the Hyper-Wiener Index of Trees. *J. Chem. Inf. Comput. Sci.* **1994**, 34 , 1079-1081.

FIGURE CAPTIONS

Figure 1. Plot of W'/W *vs* R·10^{-2} for nonanes. The regression equation and statistical parameters are W'/W = -1.055(±0.011)(R·10^{-2}) + 6.899(±0.023); n = 35; r^2 = 0.997; s = 0.026; $F^{1,33}$ = 9507.

Figure 2. a) A connected undirected acyclic graph T. An edge of T is denoted by e_{ij}.
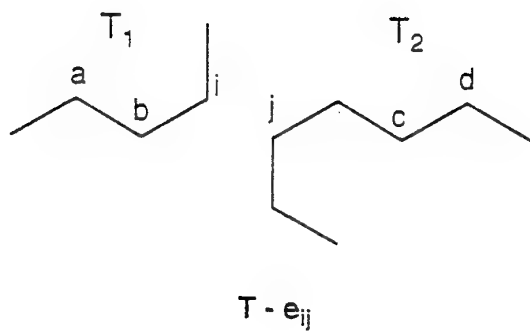b) The spanning subgraph T-e_{ij} of T with components T_1 and T_2.

Figure 3. Calculation of the R index of the hydrogen suppressed graph of 2,2,3,4-tetramethylpentane. Numbers at each site represent the corresponding graph-theoretical valences.

a)

T

b)

T_1 T_2

$T - e_{ij}$
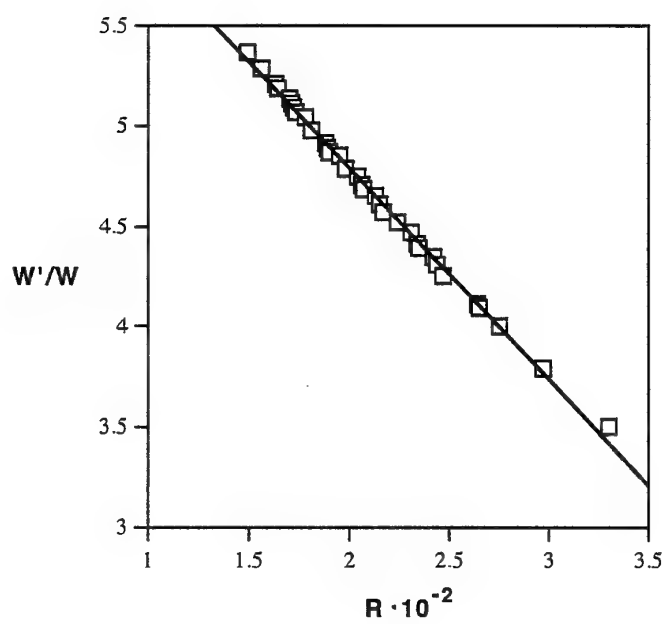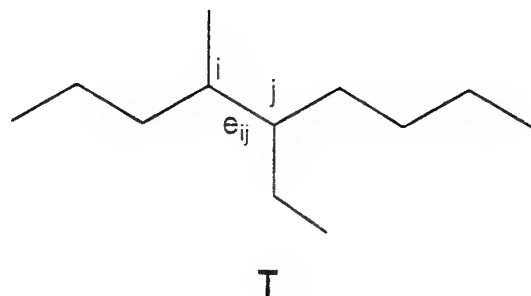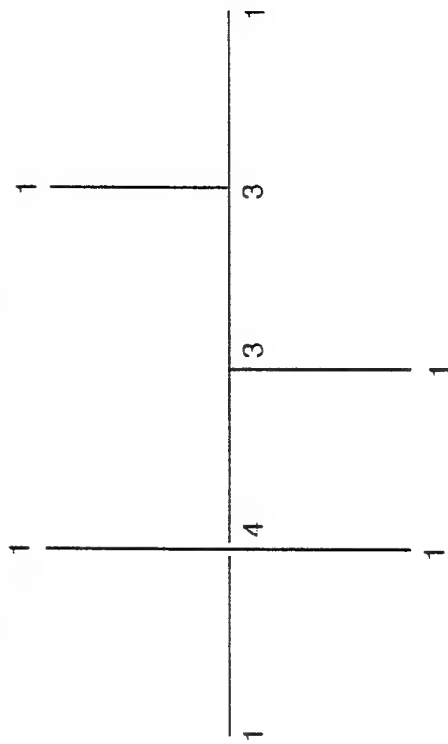
$$R(T) = \frac{1}{2}\left\{ 2\times8 + 6\left| \frac{1}{2}\left(6\times1^2 + 2\times3^2 + 1\times4^2\right) - 8 \right| + 12\left[3\times1\times4 + 3\times1\times3 + 3\times3 + 3\times4 - \left(6\times1^2 + 2\times3^2 + 1\times4^2\right) + 8 \right| + \right.$$

$$\left. 20\left| -2\left(3\times1\times4 + 3\times1\times3 + 3\times3 + 3\times4\right) + 4\times1\times1 + 6\times1\times3 + 3\times4 + 4\times1 + \frac{3}{2}\left(6\times1^2 + 2\times3^2 + 1\times4^2\right) - 8 \right| \right] \right\} = 164$$

Dynamical Process of Exciton Combination in Polymers *

E. H. Zhao, H. Jiang, X. H. Xu and X. Sun

Department of Physics, Fudan University, and National Laboratory of Infrared Physics

Shanghai 200083, *China*

Thomas F. George

Office of the Chancellor / Departments of Chemistry and Physics & Astronomy

University of Wisconsin − Stevens Point, Stevens Point, Wisconsin 54481 − 3897, *USA*

Abstract

Due to the one-dimensional characteristic, the excitation of an electron in polymers is self-trapped, and two excitons can be combined to form a biexciton. This combination process is an important channel to forming the biexciton and is accompanied by lattice distortion. By solving the dynamic equations, the relaxation process of this combination is investigated, where it is found that the relaxation time is about 160 fs.

I. INTRODUCTION

Conjugated polymers with nondegenerate ground states, such as polyparaphenylenevinylene (PPV) and polyphenylquinoxaline (PPQ), are found to have excellent nonlinear optical properties and can be used as the active luminescent layer in new polymer-based electroluminescence LED devices [1]. The excited states, especially excitons and biexcitons play central roles in these photophysics processes [2]. Recently it was proposed that the biexciton state possesses a novel property — negative polarizability, where its induced dipole moment is in the opposite direction to the external electrical field [3]. This is a straightforward way for the exciton to absorb a photon and become a biexciton, but the efficiency of this two-photon process is low. There is another channel to form the biexciton. The polymer chain is a quasi-one-dimensional system, and due to strong coupling between electron and lattice motion, the electron excitation must be accompanied with bond structure distortion. This self-trapping effect increases the binding energy of the excited state [4]. As a result, the energy of a self-trapped biexciton is lower than that of two separate excitons. Thus, two self-trapped excitons in one chain can evolve to a single biexciton. Under photoexcitation, many excitons are generated, and in the case of electroluminescence, the injection of charge carriers also produces many excitons. When two moving excitons encounter each other, they combine into a biexciton, which is an important channel for biexciton formation [5].

In this paper, the evolution of the combination is simulated by solving the dynamic equations. From the evolution of the bond structure and electronic spectrum, it is found that the

2

relaxation time of the combination is about 160 fs.

II. MODEL AND METHOD

As usual, the SSH Hamiltonian [4] (modified by Brazovskii and Kirova [6]) is used to model the polymer in its nondegenerate ground state, consisting of N lattice sites and N electrons:

$$H_0 = \sum_n [t_0 - \alpha (u_{n+1} - u_n) + (-1)^n t_e] \left(a_{n+1,s}^+ a_{n,s} + h.c. \right)$$

$$+ \frac{K}{2} \sum_n (u_{n+1} - u_n)^2 - K' \sum_n (u_{n+1} - u_n) \ . \tag{2.1}$$

Here u_n is the displacement of the nth atom, $a_{n,s}$ is the annihilation operator of an electron at the nth atom, the nondegenerate ground state is depicted by t_e, and the final term is added to avoid the collapse of the open chain with a finite length [7,8]. When an external electrical field E is applied, the potential energy of electrons is

$$H_1 = eE \sum_{n,s} \left(na - \frac{N+1}{2} a + u_n \right) a_{n,s}^+ a_{n,s} \ . \tag{2.2}$$

We define the order parameter as $\phi_n = (-1)^n u_n$ and divide the total Hamiltonian into an electron part H_e and a lattice part,

$$H (\{\phi_n\}) = H_0 + H_1 = H_e + \frac{K}{2} \sum_n (\phi_{n+1} + \phi_n)^2 - K' \left(\phi_1 + (-1)^N \phi_N \right) \ . \tag{2.3}$$

For a certain lattice configuration $\{\phi_n\}$, we can diagonalize $H_e(\{\phi_n\})$ to obtain the electronic energy spectrum $\{\varepsilon_i\}$ and corresponding eigenstate $|i\rangle$. Then the total energy is

3

$$E\left(\{\phi_n\}\right) = \sum_i^{occu} \varepsilon_i + \frac{K}{2} \sum_n \left(\phi_{n+1} + \phi_n\right)^2 - K'\left(\phi_1 + (-1)^N \phi_N\right). \tag{2.4}$$

Because the mass of an atom is much larger than that of an electron, we have been able to use the adiabatic approximation [9]. The nth atom experiences a force

$$f_n = m\frac{d^2\phi_n}{dt^2} = -\frac{\partial E\left(\{\phi_n\}\right)}{\partial \phi_n} = \sum_i^{occu} \frac{\partial \varepsilon_i}{\partial \phi_n} + K\left(\phi_{n+1} + 2\phi_n + \phi_{n-1}\right) - K'(\delta_{n1} + (-1)^N \delta_{nN}).$$

$$\tag{2.5}$$

By using the Hellman-Feynman theorem, we get

$$\frac{\partial \varepsilon_i}{\partial \phi_n} = \sum_i (-1)^n \langle n|i\rangle \left(2\alpha \langle n+1|i\rangle \left(1 - \delta_{n,N}\right) - 2\alpha \langle n-1|i\rangle \left(1 - \delta_{n,1}\right) + \langle n|i\rangle eE\right). \tag{2.6}$$

Taking a short interval τ as the time step, the dynamic equations (2.5) can be solved numerically. Step by step, we can simulate the dynamic evolution of the bond structure. The period τ_0 of lattice vibration is about 4×10^{-14} s, and since the time step τ must satisfy $\tau \ll \tau_0$, we choose $\tau = 1$ fs. The damping term is $-\lambda m d\phi/dt$, and λ should be $\ll 1/\tau_0$ [10,11]. The results show that changes in λ do not influence the main character of the relaxation process. In our calculation, we have chosen the parameters according to the cis-polyacetylene,

$$t_0 = 2.5 \ eV, \quad t_e = -0.05 \ eV, \quad \alpha = 41 \ eV/nm, \quad K' = 1.25 \ \alpha$$

$$K = 2.1 \times 10^3 \ eV/nm^2, \quad a = 0.122 \ nm, \quad m = 13 \ u, \quad N = 100 \ .$$

III. RESULTS

To simulate the combination of two excitons into one biexciton, we take two excitons which are close to each other as the initial lattice configuration ($t = 0$ in Fig. 2). For an exciton, an electron is excited to the bottom of the conduction band (LUMO). For an biexciton, two electrons with opposite spins are excited to the LUMO with two holes left in the HOMO. Figure 1 shows the evolution of the 50th energy level E_{50} under different electrical fields. At $t = 0$, E_{50} is the energy level of the lower gap state of the self-trapped exciton. As time passes, it evolves into that of the self-trapped biexciton. Figure 2 shows the lattice configuration at different times for $E = 0.1$ Mv/cm. As can be seen from these figures, when $t > 100$ fs , the change of lattice configuration becomes smaller and smaller, and finally converges.

From Figs. 1 and 2 it is seen that the relaxation time is about 160 fs, which is about the same as that of the photoexcitation process. As a result of exciton combination, the energy of the system decreases, and the binding energy of the biexciton is 0.77 eV.

REFERENCES

[1] J. H. Burroughes, D. D. Bradley, A. R. Brown, R. N. Marks, K. Mackay, R. H. Friend, P. L. Burn and A. B. Holmes, *Nature* **347**, 539 (1990).

[2] V. A. Shakin, *Phys. Rev. B* **50**, 4306 (1994).

[3] R. L. Fu, H. J. Ye, L. Li and X. Sun, *Acta Physica Sinica* **47**, 94 (1998).

[4] A. Heeger, S. Kivelson, J. R. Schrieffer and W. P. Su, *Rev. Mod. Phys.* **60**, 821 (1988).

[5] D. W. McBranch, R. Kohlman, V. I. Klimov, B. Kraabel, A. J. Heeger and B. Hsieh, *Abstracts of the International Conference on Science and Technology of Synthetic Metals* (Montpellier, France, 1998), page 93, TUC86.

[6] S. A. Brazovskii and N. N. Kirova, *JETP Lett.* **33**, 4 (1981).

[7] S. Kivelson, T. K. Lee, Y. R. Lin-Liu, I. Peschel and L. Yu, *Phys. Rev. B* **25**, 4173 (1982).

[8] W. P. Su, *Solid State Commun.* **35**, 899 (1980).

[9] W. P. Su and J. R. Schrieffer, *Proc. Nat. Acad. Sci. (USA)* **77**, 5626 (1980).

[10] X. Sun, G. P. Zhang, Y. S. Ma, R. L. Fu, X. C. Shen, K. H. Lee, T. Y. Park, T. F. George and L. N. Pandey, *Phys. Rev. B* **53**, 15481 (1996).

[11] X. Rao, R. T. Fu, X. Sun and K. Nasu, *Phys. Lett. A* **266**, 383 (1997).

Figure Captions

Fig. 1. Evolution of the energy level E_{50} under different electrical fields: $E = 0$ and $E = 0.1 \ Mv/cm$.

Fig. 2. Evolution of the lattice configuration under $E = 0.1 \ Mv/cm$: (a) t from 0 to 60 fs; (b) t from 80 to 100 fs; (c) t from 120 to 180 fs.
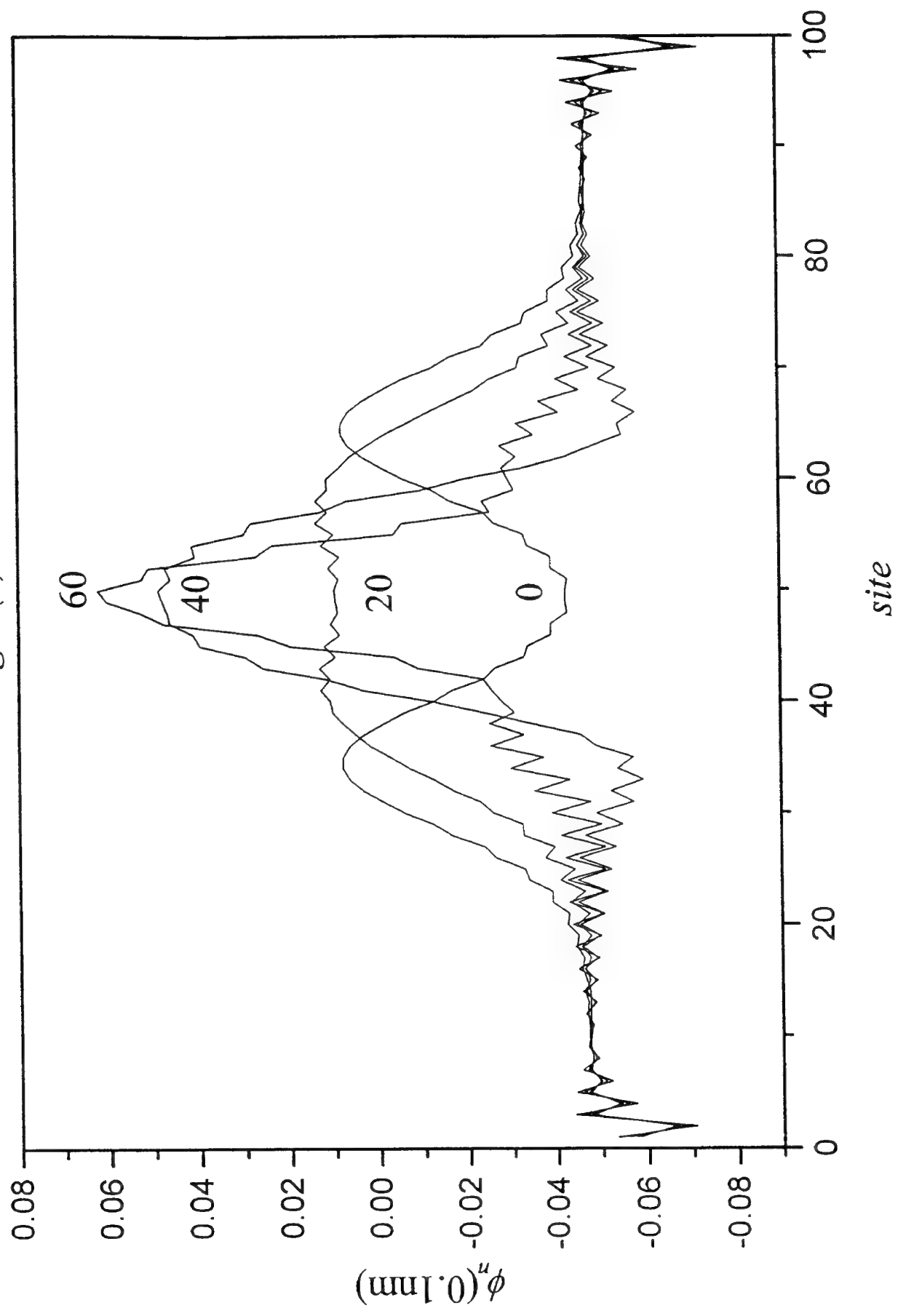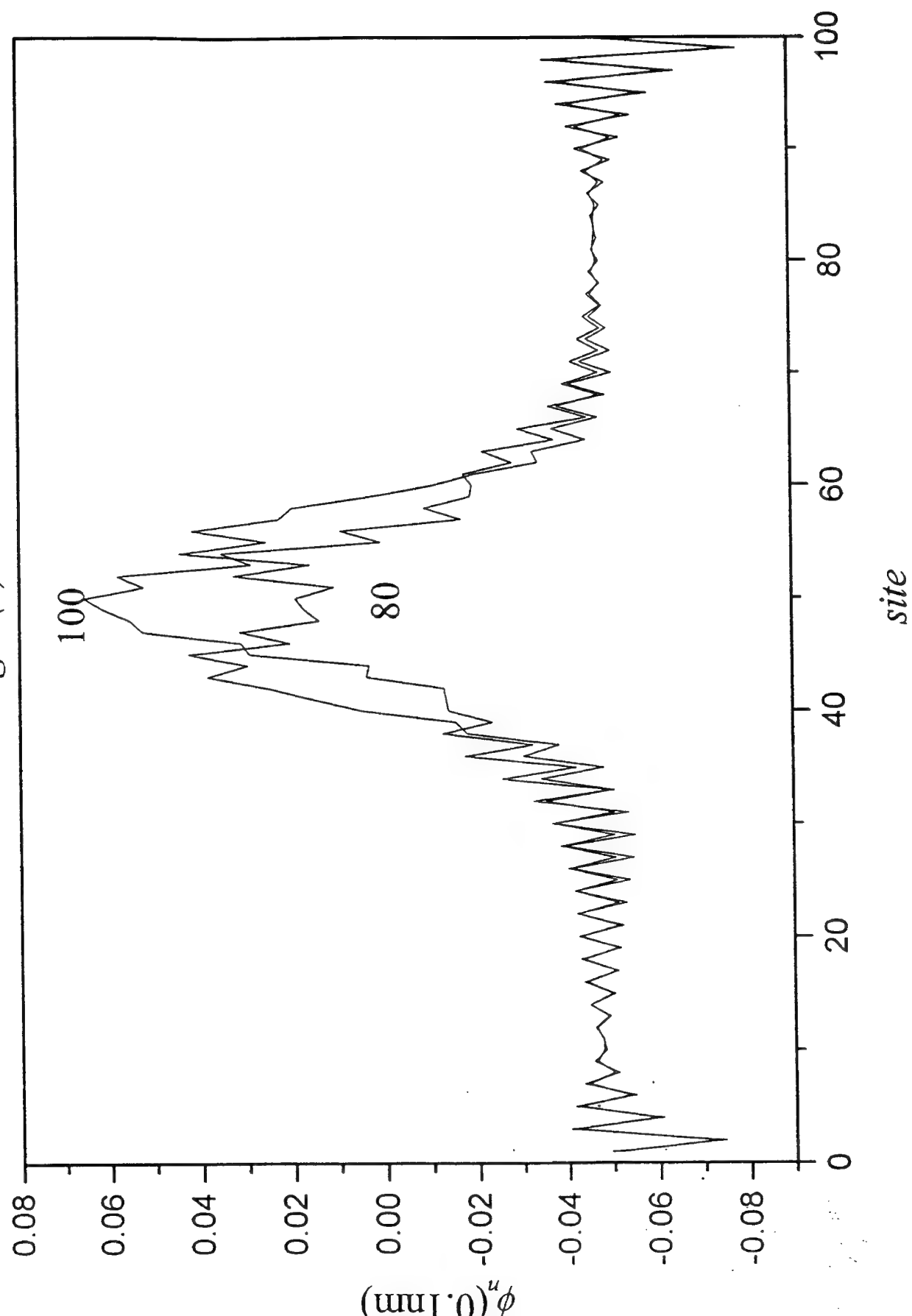
Fig.1

The 50th energy level (eV)

time (fs)

E=0

E=0.1 MV/cm
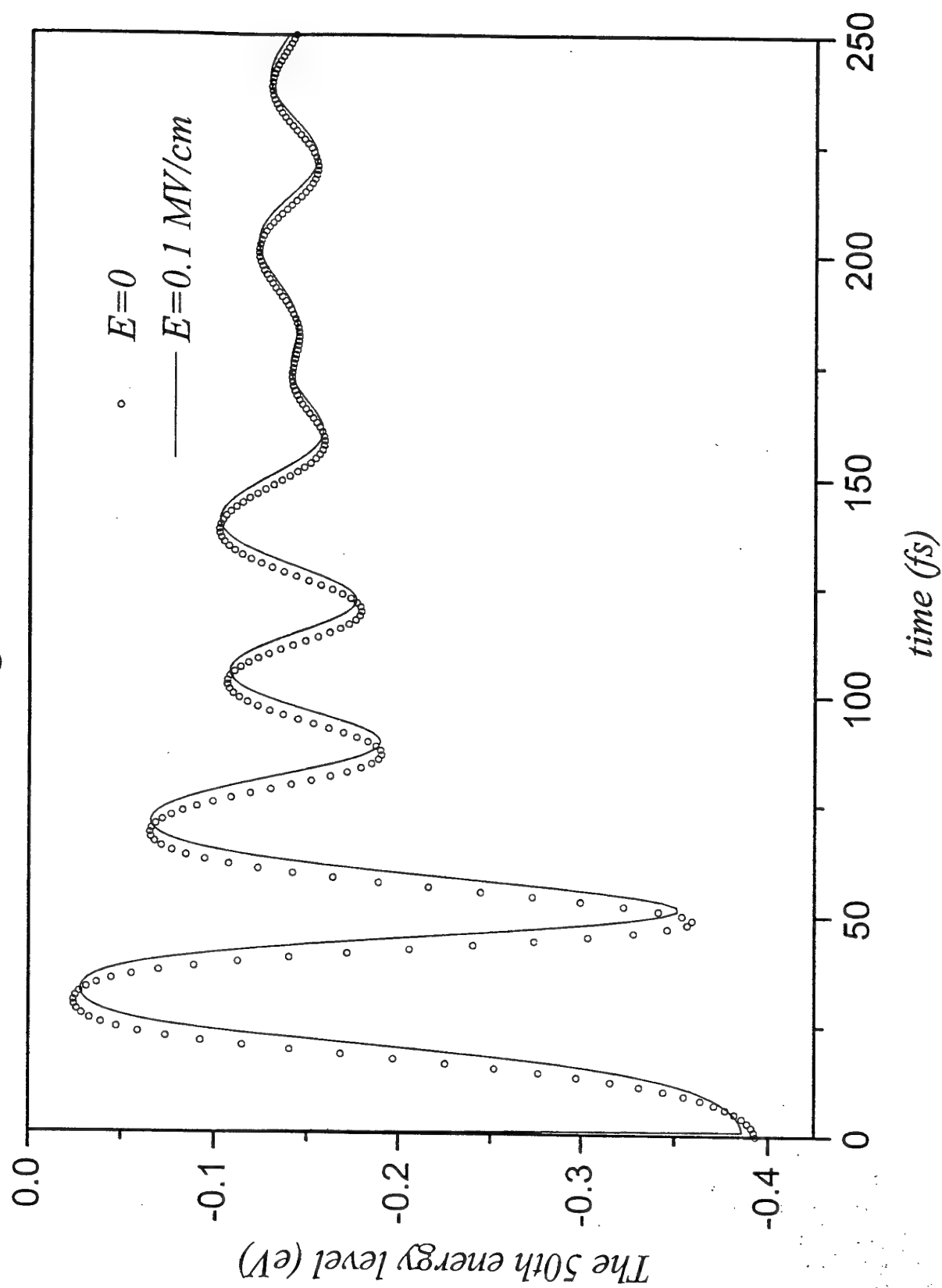
Fig.2 (a)

Fig.2 (b)

Fig.2 (c)
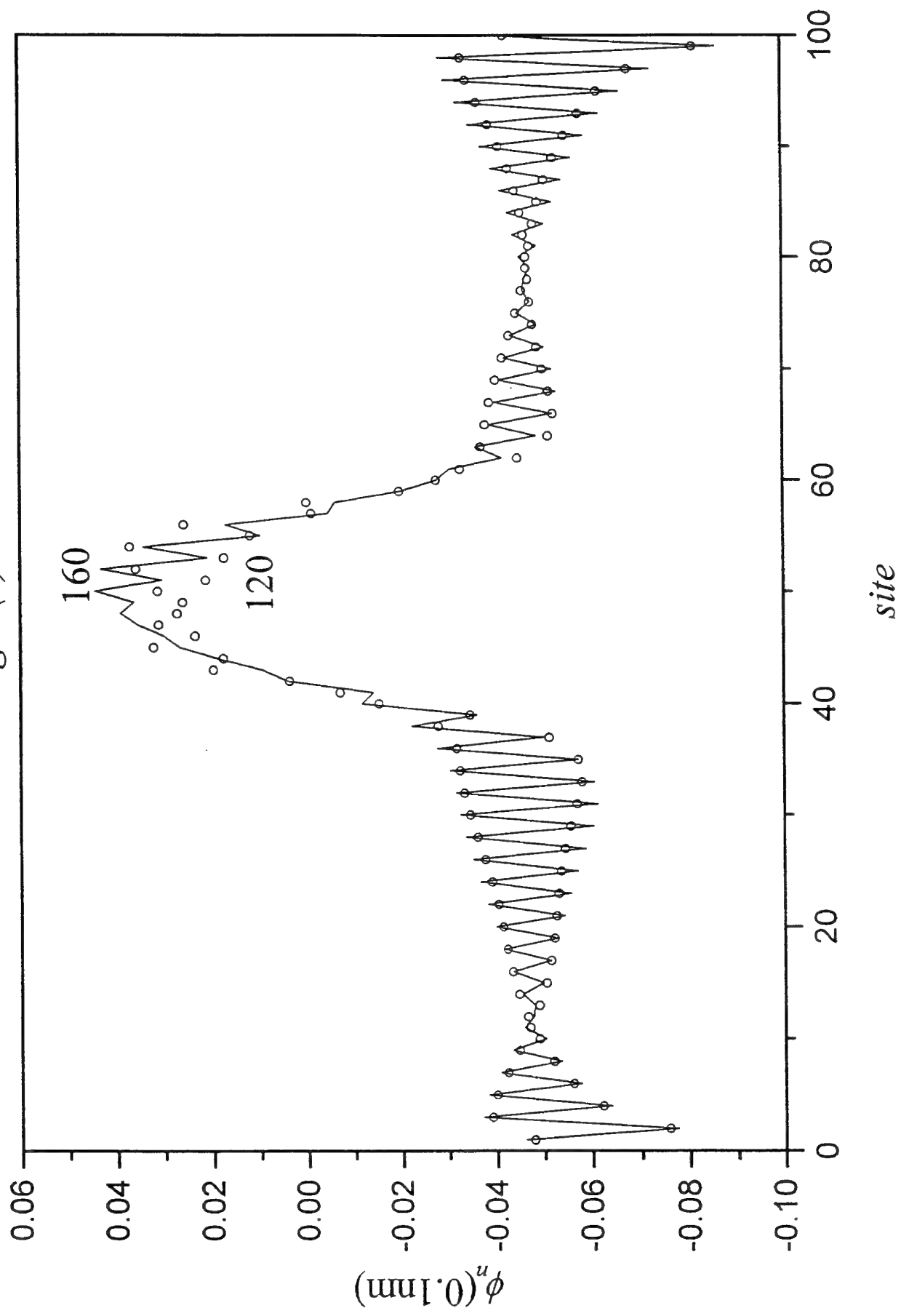
PRECIPITATION AT EQUIVALENCE AND EQUILIBRIUM - A METHOD FOR THE DETERMINATION OF EQUILIBRIUM CONSTANTS OF REACTION BETWEEN MULTIDETERMINANT ANTIGEN AND SPECIFIC POLYCLONAL ANTIBODIES

Biserka Pokrić

Ruđer Bošković Institute, Zagreb, POB 1016, Croatia

e-mail: pokric@rudjer.irb.hr

A theoretical approach for the determination of the equilibrium constant, K_a, of the reaction between a multideterminant antigen (Ag) and specific polyclonal antibodies (Ab) forming the insoluble Ab/Ag immune complex, is derived. The constant can be expressed as a function of the two accessible experimental parameters, the precipitating concentration of the antigen and the Ab/Ag molar ratio. For this purpose Ab/Ag immune complex must be prepared at equivalence and equilibrium between precipitated and soluble species must be reached. The proposed method is experimentally tested on the system human serum albumin (HSA) and polyclonal rabbit antibodies. The Ab/Ag precipitates are prepared by the direct mixing of biological fluids in which immunoreacting components naturally occur. Previous separation, purification or labelling of immunoreacting components are not required. The conditions for the precipitation of Ab/Ag complexes at equivalence, the stoichiometric composition or the average number of Ab molecules bound to one Ag molecule, and the solubility of the immunoprecipitating components, are determined by a rectangular two-dimensional double immunodiffusion. Since the solubility determined under the conditions of a double immunodiffusion is a result of the interaction of the global diffusion of the precipitating components and particle growth kinetics, it mostly refers to the dynamic conditions. In order to find the solubility under equilibrium conditions, it is sufficient to determine the minimal factor by which the solutions of both immunoprecipitating components should be diluted so that no precipitate is formed upon their mixing at equivalence. The dilution factor is determined by a measurement of the laser light scattering of the immunocrecipitating systems prepared with serially diluted Ag and Ab solutions.

INTRODUCTION

Affinity, specificity and concentration of an antibody (Ab) determine its usefulness for analytical, diagnostic and therapeutic purposes and are also implicated as important factors in the immune response. Antibody affinity is defined as the attractive force between an antigenic determinant (epitope) and the antibody combining site (paratope). Accordingly to its definition, the affinity can be measured only when the antigen (Ag) is a simple, well-defined substance such as a hapten. The determination of the affinity of antibodies directed against a protein in this sense is impossible, because of the multiplicity and heterogeneity of antigenic determinants. To describe quantitative differences of the interactions between a multideterminant Ag and polyclonal Abs, despite a lack of a precise thermodynamic and immunochemical meaning, the term avidity was introduced. Avidity is the goodness of fit between more than one epitope of the antigen and more than one site of the antibody. The rates of dissociation, the solubilities of Ab/Ag complexes precipitated at the optimum proportions, the deviations from the linearity of the curve representing reciprocal concentrations of a bound antigen vs. reciprocal concentrations of a free antigen, antigen binding capacities, immunochemical titers, indices of avidities, as well as the data obtained by measurements of 50% binding of antibodies by an antigen are related to the avidity.[1-3] However, the avidities are often expressed through the association constants, K_a, which represent the equilibrium of association, k_a, and dissociation, k_d, rates of Ab/Ag complexes. The difficulties in reaching equilibrium due to heterogeneous binding are responsible that the application of the solid phase affinity methods for the determination of K_a, should be taken with precautions. Despite of the simplicity of these methods, such as an ELISA, various surface effects can cause errors in estimates of either liquid or solid phase affinities and influence on the ranking of affinities.[2,4,5]

In this paper we propose a new approach to the determination of the equilibrium constants of the reactions between multideterminant Ags and specific polyclonal Abs. The method is based solely on the determinations of the concentration of soluble Ag in equilibrium with insoluble Ab/Ag immune complex prepared at equivalence.

THEORETICAL CONSIDERATIONS

Equilibrium Constants of Polyclonal Antibodies Raised Against a Multideterminant Antigen. An immune complex of the average composition AB_n is formed when one molecule of multideterminant antigen A reacts with n molecules of the specific polyclonal antibodies B:

$$A - n B \leftrightarrow ABn$$

The average composition means that n should not necessary be an integer number. In case that n is greater than unity, a portion of the immune complex ABn may be precipitated. In this case the equilibrium is reached when fully reversible reaction is established between the precipitated complex ABn_p (solid phase) and the species remaining in the solution (solute phase):

$$ABn_p \leftrightarrow ABn_s + A_s + n\,B_s$$

The solute phase consists of the free components A_s and B_s, and components $A_{s,b}$ and $B_{s,b}$ bound in the soluble ABn_s complex. The reversible reaction between soluble immune complex and free species in solute phase, also exists under equilibrium conditions :

$$ABn_s \leftrightarrow A_s + n\,B_s \tag{1}$$

In the immunoprecipitating system prepared under equivalence conditions, the ratio of molar concentrations of B and A components, n, is:

$$n = cB/cA \tag{2}$$

and is the same in the precipitated and soluble immune complexes and corresponds to the ratio of free B and A species in the solute phase.

According to eq 1, the equilibrium constant of the primary Ab/Ag reaction in a system prepared at equivalence and equilibrium is:

$$Ka' = cA_{s,b}/(cA_s\,cB_s^{n}) \tag{3}$$

By substitution of the cB by cA from eq 2, the equilibrium constant equation becomes:

$$Ka' = cA_{s,b}/(n^{n}\,cA_s^{n+1}) \tag{4}$$

Ka' reflects the avidity of the antibodies, irrespective of the number of antibody valences, i.e, binding sites, involved in the reaction. In case that the multideterminant antigen bears distinct antigenic determinants, each of them is able to react only with one binding site on the molecule of the specific antibody. Since the antibody molecules possess more than one binding site, the valency of antibodies should be taken into account in order to calculate the equilibrium constant. For instance, the equilibrium constant of the reaction involving bivalent antibodies belonging to IgG classes, reads as:

$$\ln Ka = \ln Ka'/2\,n \tag{5}$$

Equivalence and Equilibrium Conditions. In case that the concentrations of the antigen and antibody solutions are unknown, the equivalence conditions can be determined by a rectangular two-dimensional double diffusion technique called the "two-cross" immunodiffusion. A detailed description of the two-cross experimental set-up which enables an adequate solution of the Fick's second law of diffusion applied to the immunoprecipitation in gels, is described elsewhere.[6,7] Briefly, a "cross" consists of four troughs cut at a right angle in a gel plate. The half-width of the trough is denoted by h. The troughs of each cross are filled in alternate order with antigen solution (component A) and immune serum (component B). In the second cross, the solutions of both precipitating components are diluted by the same factor, d. The distances between peaks of the precipitin lines in the direction of the diffusion of antigen, x, and antibodies, y, are measured in both crosses at a same time, t. The volume ratio of the solutions of the precipitating components required to ensure equivalence conditions during the preparation and precipitation of immune complexes[8], is given by:

$$vB/vA = ad^b \tag{6}$$

where

$$a = [(x_1^2 - x_2^2)/(y_1^2 - y_2^2)]^{1/2}$$

$$b = (x_2^2 y_1^2 - x_1^2 y_2^2)/[(x_1^2 - x_2^2)(y_1^2 - y_2^2)]$$

The subscript 1 and 2 refer to the parameters measured in the first and the second cross, respectively.

So far, the two-cross immunodiffusion technique enables the direct determination of the reciprocal precipitating titers (eqs 7 and 8) and the diffusion coefficients (eqs 9 and 10) of the reacting molecules. Precipitating titers PT are defined as the ratio of the equivalent molar concentrations of the substance at the origin of diffusion, c_0, and at the point of the onset of precipitation, c_{pr}:

$$1/PTA = (W/x_1) X e^{-X} = cA_{pr}/cA_0 \tag{7}$$

$$1/PTB = (W/y_1) Y e^{-Y} = cB_{pr}/cB_0 \tag{8}$$

where:

$$W = h (2/\pi)^{1/2}$$

$$X = x_1 [2 \ln d/(x_1^2 - x_2^2)]^{1/2}$$

$$Y = y_1 [2 \ln d/(y_1^2 - y_2^2)]^{1/2}$$

The precipitating concentrations, c_{pr}, can be calculated from known both precipitating titers and initial concentrations of the solution of the precipitating components. In case that the initial concentrations of the antibody and antigen solutions are unknown, they can be determined directly in crude biological fluids in a manner described in details elsewhere.[8,9]

The diffusion coefficients are obtained from the relations:

$$DA = 1/t [(x_1^2-x_2^2)/(4 \ln d) - h^2/6] \tag{9}$$

$$DB = 1/t [(y_1^2-y_2^2)/(4 \ln d) - h^2/6] \tag{10}$$

From the data obtained for the diffusion coefficients, the approximate values of the molecular masses of the Ag and Ab molecule can be calculated using a simple relation:

$$M = M_{IgG} (D_{IgG}/D)^3 \tag{11}$$

$M_{IgG} = 150\ 000$ Da and $D_{IgG} = 4.1\ 10^{-7}$ cm^2/s is the molecular mass and the diffusion coefficient of human IgG, respectively. D is the diffusion coefficient of Ag or Ab molecule referring to the free diffusion in distilled water at 20 °C.

From the conservation of mass, the precipitating concentrations referring to the equilibrium conditions correspond to the sum of the concentrations of free reactant and the concentration of reactant bound in a soluble immune complex:

$$cA_{pr} = cA_{s,b} + cA_s \tag{12}$$

$$cB_{pr}r = cB_{s,b} + cB_s \tag{13}$$

According to the equivalence rule, the precipitation under conditions of a double diffusion starts at the equivalent molar concentrations of both precipitating components,[7,10-14] as refers eq 2:

$$n = cB_{pr}/cA_{pr} \tag{14}$$

This means that the equilibrium constant equation (eq 4) can be solved using c_{pr} values from eq 7 or 8. The critical precipitating concentrations in eqs 7 and 8 represent the solubility of an immunoprecipitating component under dynamic conditions, as a result of the interaction of the global diffusion of the precipitating components and particle growth kinetics.[10] Thus, the solubility, c_{pr}, has a kinetic and not a thermodynamic significance. In case when enough time is allowed for the precipitation in the solutions to reach the equilibrium, the solubility referring to the thermodynamic conditions, could be found. If not, it is sufficient to determine the minimal factor, m_0, by which both solutions of A and B components should be diluted so that no precipitate is formed upon their mixing at equivalence. The concentrations, c_{pr}/m_0, referring to the equilibrium conditions are denoted as $c_{pr}*$.

In order to solve eq 4 using $cA_{pr}*$ and/or $cB_{pr}*$, the concentrations of free components and components bound in an immune complex in the solute phase (eqs 12 and 13) should be known. The solution is possible by introducing the ratio, r, of the concentrations of both bound and free antigen in the solute phase:

$$r = cA_{s,b}*/cA_s*$$

The equilibrium concentrations cA_s* and $cA_{s,b}*$, according to the definition for cA_s and $cA_{s,b}$ given by eq 12, can be now expressed as follows:

$$cA_s* = cA_{pr}*/(r + 1) \tag{15}$$

$$cA_{s,b}* = cA_{pr}* \, r/(r + 1) \tag{16}$$

The equilibrium constant equation (eq 4) in terms of $cA_{pr}*$, r and n, reads as:

$$Ka' = (1/cA_{pr}*)^n \, r \, [(r + 1)/r]^n \tag{17}$$

Equation 17 consists of two factors: the factor $(1/cA_{pr}*)^n$ and a factor F which is a function of r and n:

$$F = r \, [(r + 1)/n]^n \tag{18}$$

The molar concentrations of both bound and free antigen in the solute phase are extremely low and their experimental determination as well as the determination of

their ratio, r, is difficult. By varying the values of n in eq 18, the F function reaches a maximum for a constant r value (Figure 1). The maximum of F function corresponds to the maximal stability of an immunoprecipitating system prepared at equivalence under given experimental conditions (pH, temperature, ionic strength, etc.) and now, the Ka' equation reeds as:

$$Ka' = (1/cA_{pr}^*)^n \, F_{max} \tag{19}$$

F_{max} is solely a function of n. In order to find F_{max}, the derivative of F should be zero:

$$dF/dn = (r/n^2) \, [(r + 1)^n \ln (r + 1) - (r + 1)^n (\ln n + 1)] = 0$$

this means that:

$$\ln (r + 1) = \ln n + 1 = \ln (n \, e)$$

i.e.

$$r = n \, e - 1 \tag{20}$$

Introducing eq 20 into eq 18

$$F_{max} = (n \, e - 1) \, e^n \tag{21}$$

where e is the basis of the natural logarithms.

The solution of the equilibrium constant eqs 4 and/or 17 in terms of cA_{pr}^* and n is:

$$Ka' = (e/cA_{pr}^*)^n (n \, e - 1) \tag{22}$$

MATERIALS AND METHODS

Immunogen and Antigen. Immunochemically pure human serum albumin (HSA), pI 4.7, M.w. 65 kDa was prepared under non-denaturating conditions and contained 97.6% of monomer and 2.4% of dimer (Calbiochem-Behring Corp., La Jolla, Ca., USA). The same HSA was used for the immunization of rabbits (immunogen) and as the antigen in the precipitation experiments. The concentration of antigen solutions amounted to cA0 = 0.500 g HSA/l.

Animals, Immunization and Immune Sera. Three randomly chosen male New Zeland rabbits were immunized successively in two-week intervals with the HSA. Each dose of the immunogen contained 0.75 mg of the HSA in a volume of 1 ml of complete (CFA) or incomplete Freund's adjuvant (IFA) or saline solution. For the primary immunization, the immunogen was emulsified in the CFA and administered i.d. in the region of the peritoneal cavity. For the first and the second booster doses the HSA was emulsified in the IFA and administered i.d in the region of the peritoneal cavity. For the third booster dose a solution of HSA in saline was applied i.v. in ear veins.

Two weeks after receiving the last booster dose, rabbits were bled and blood samples collected separately. The immune sera were decomplemented at 56 °C for 30

min. The antibody concentrations of the immune sera were previously determined by a microgravimetric method.[8]

The Two-Cross Immunodiffusion Experiments. For the two-cross immunodiffusion experiments 1% w/v agarose gel was prepared using agarose L (Behring Institute, W. Germany). Phosphate buffered saline (PBS), pH 5.0, 5.5. and 7.0, contained 0.05 M KH_2PO_4, 0.10 M NaCl, 0.1% w/v NaN_3 and a variable amount of NaOH. The borate buffered saline (BBS) contained 0.05 M H_3BO_3, 0.1 M NaCl, 0.1% w/v NaN_3 and NaOH was added in order to reach pH 8.6. These buffer solutions were used to dilute the precipitating components and to equilibrate the agarose in which the two-cross immunodiffusion experiments were performed. The immunodiffusion experiments were carried out at 20 and 40 °C. For all technical details concerning the two-cross immunodiffusion experimental procedure and the evaluation of the results, refer to Pokrić and Pučar[6] or Živković et al.[7] for somewhat-less-detailed descriptions.

The Determination of the Solubility of Immune Complexes. The solubility of the immune complexes was determined by consecutive dilutions of the solutions of the precipitating components until the dilution, m_0, at which no precipitate is formed.[8] The starting precipitating system was prepared at equivalence by mixing 20 μl of the immune serum and antigen solution which volume was calculated by eq 6. For the dilutions, the total volume of the precipitating system was maintained constant, but the volume parts of the solutions of the precipitating components were subsequently reduced and simultaneously the volumes of the buffer solution increased. Relative laser light scattering (%RLLS) measurements were performed using a Hyland laser nephelometer PDQTM (Travenol Laboratories, Costa Mesa, Ca, USA). For each dilution, the measurement was carried out until the maximum values of scattered light, $(\%RLLS)_{max}$ were reached. The experiments were performed at 20 and 40 °C.

RESULTS

Figure 1 illustrates the changes of the factor F (eq 18) vs. n for certain values of r. The molar Ab/Ag ratio, n, refers to the immune complex prepared at equivalence. The ratio of the molar concentrations of bound Ag over free Ag, $r = cA_{s,b}^*/cA_s^*$, is related to the equilibrium reached in the solute phase. Figure 1 shows that for each of the chosen r values ranging between 5 and 12.5 and varying n values from 1 to 7, the factor F reaches a maximum. The n values must lay in a limited range. The values of n < 2 will seldom give precipitates at equivalence. The values of n > 5 will exceptionally occur and then the maxima of the F function vs. n (Figure 1) have a very steep shape. The latter case would drastically increase the equilibrium constant (eq 19). For n values between 1.5 and 5.5, and according to eq 21, the values of the F_{max} will lay between 13.79 and 3412. From eq 18 proceedes that in this case the molar ratios, r, of associated, $cA_{s,b}$, and free antigen, cA_s, in the solute phase will predominantly lay between 3.07 and 13.95.

The diffusion coefficients of antigen and antibody, DA = 6.1 x 10^{-7} cm^2/s and DB = 4.1 x 10^{-7} cm^2/s, determined by the immunodiffusion method (eqs 9 and 10), correspond to the HSA and IgG class rabbit antibodies having molecular mass MA = 65 000 Da and MB = 160 000 Da (eq 11), respectively.

An example of the determination of the dilution factor, m_0, at which no detectable precipitate is formed, is presented in Figure 2. The quantity of the precipitate formed

was determined by the laser light scattering measurements. A strict linearity, $r > 0.9$, between the intensity of scattered light, $(\%RLSS)_{max}$ and the dilution, $1/m$, was obtained in all examined systems. Thus, it is not necessary to find experimentally the dilution at which precipitation occurs no more. The extrapolation of the straight line $(\%RLSS)_{max}$ vs. $1/m$ to $\%RLLS = 0$ gives the required dilution $1/m_0$.

The precipitating titers of both reactants determined by the two-cross immunodiffusion (eqs 7 and 8) and the dilutions factors required for the corrections to equilibrium conditions, are presented in Table 1. The values obtained for the dilution factor, $m > 1$, proved that the precipitating titers determined under conditions of the double diffusion do not refer to equilibrium conditions. Table 1 and Figure 2 shows that m_0 is always smaller at 40 than at 20 °C. This suggests that the precipitation under conditions of diffusion at higher temperatures, occurs under conditions closer to that of equilibrium.

The precipitating concentrations, c_{pr}, of antigen and antibodies at which the precipitation starts under conditions of double diffusion (eqs 7 and 8), are presented in Table 2. Taking into account antibody and antigen molecular masses, the molar Ab/Ag ratio, n, required for the formation of the immune complex at equivalence, was calculated (eq 14). The stoichiometric composition of an Ab/Ag complex prepared under identical experimental conditions but at two different temperatures, is constant (Table 2). The small differences between, n, at 20 and 40 °C arise from the errors in experimental determinations of the precipitating concentrations, cA_{pr} and cB_{pr}, (eqs 7 and 8).

The mean values of n are used in order to calculate the equilibrium constants Ka' (eqs 22). In order to calculate Ka', the precipiting concentrations of antigen solutions obtained in g/l (Table 2) must also be expressed in mol/l and corrected to equilibrium conditions, cA_{pr}^* by using dilution factors, m_0 (Table 1). Assuming that IgG class Abs are bivalent, the equilibrium constants, Ka, are calculated according to the eq 5, and presented in Table 3. The precipitating titers of the antigen solution corrected to equilibrium conditions, PTA* are also presented in Table 3. In spite of the fact that in our experiments the same antigen solution of a constant concentration, cA_0, was used, different PTA* values were obtained under different experimental conditions. However, the data presented in Table 3 shows that Ka and PTA* values are well-related. Thus, the PTA* data could be used for the rough ranking of the avidities of different immune sera for the same multideterminant antigen.

Lower Ka values (Table 3) at higher rather than at lower temperatures, suggest that antiHSA/HSA binding is an exothermic process.

DISCUSSION

The method, proposed for the determination of the equilibrium constants of the reaction of the mutideterminant antigen and the specific polyclonal antibodies, requires only the knoweledge of two accessible experimental parameters: the concentration at which the antigen starts to precipitate under equilibrium conditions and the molar Ab/Ag ratio in the immune complex prepared at equivalence. The immune complexes can be prepared by the direct mixing of biological fluids in which immunoreacting components naturally occur. This offers a great advantage of dealing with unmodified molecules, since the separation, purification and labelling of either the antigen or the antibody, which might modify the binding properties,[15] is not required.

Different values of precipitating titers of the antigen solution corrected to equilibrium conditions, PTA*, (Table 3) obtained for the same Ab/Ag system at different pH values, as well as under the identical experimental condition at two different temperatures (Table 1), proved that the solubility, cA_{pr} (eq 8), of a HSA-anti HSA system is dependent on both pH and temperature. The changes of the stoichiometry of an antiHSA-HSA pair at various pH (Table 2), previously observed in the experiments with a number of Ab/Ag complexes,[8,16,17] are caused by the changes of the charge of Ab and Ag molecules with ambient pH.[18,19] The dependence of both the solubilty and stoichiometry of the Ab/Ag system regarding the experimental conditions, explain the variations of equilibrium constant values of a given antiHSA-HSA pair at various experimental conditions (Table 3) since Ka is directly related to the solubility, cA_{pr}, and Ab/Ag ratio, n, (eqs 5 and 22). This finding agrees with the literature data that the value of the equilibrium constant of an Ab/Ag system is affected by the conditions under which the determination was carried out.[1,18]

A part of the difficulties of the determination of the equilibrium constants of reaction between multideterminant antigen and polyclonal antibodies lie in the fact that Ka values are often dependent upon the absolute amounts of antigen and antibodies, the dilution and/or the volume of the immunoreacting system, as well as upon the ratio of Ab/Ag concentrations.[20-25] So far, the state of equilibrium is disturbed and the dissociation rate is greatly increased when one of the precipitating components is present in a great excess.[26] In our experiments the Ab/Ag concentration ratio is determined in advance by preparing the precipitating system at equivalence, while the Ka determined at equilibrium is invariable to the total concentrations of antigen or antibodies in biological fluids.

The determination of the equilibrium constant Ka in our experiments was possible from the data obtained by the two-dimensional double immunodiffusion concerning the preparation of an Ab/Ag system at equivalence, the precipitating titer of antibody solution, PTA, and diffusion coefficients, D, and/or molecular masses, M, of immunoreacting molecules. The concentration of antigen solution, cA_0, should be known or determined in advance in order to calculate the critical precipitating concentration of antigen, cA_{pr}, (eq 7) required for the determination of Ka' (eq 22). The determination of the precipitating titer (PT) by the two-cross immunodiffusion does not require the use of the standards as well as the knowledge of the concentrations of the solutions of the precipitating components. For a constant concentration of the antigen solution, cA_0, the precipitating titers referring to equilibrium conditions, PTA*, depend solely on the critical precipitating concentrations, $cA_{pr}*$ (eq 7). So far, according to the theory, the different concentrations of identical antibodies obtained by dilutions of an immune serum, would not influence the PTA* and/or $cA_{pr}*$ values.[8] The comparison of PTA* and Ka values (Table 3) shows that they are well-related. Thus, under the conditions when the concentrations and molecular masses of antigens and antibodies are unknown or difficult to determine, the PTA* could be used for a rough ranking of relative affinities of different immune sera against a same antigen.

REFERENCES

(1) Absolom, D.R.; van Oss, C.J. The nature of the antigen-antibody bond and the factors affecting its association and dissociation. *Crit. Rev. Immunol.* **1986**, 6, 1-46.

(2) Azimzadeh, A.; van Regenmortel, M.H.V. Antibody affinity measurements. *J. Mol. Recognit.* **1990**, 3, 108-116.

(3) Steward, M.W.; Steensgaard, J. *Antibody Affinity: Thermodynamic Aspects and Biological Significance.* CRC Press. Inc. : Boca Raton, Fl.; **1983**, pp 1-98.

(4) Loomans, E.E.M.G.; Roelen, A.J.M.; Van Damme, H.S.; Bloemers, H.P.J.; Grinbau, T.C.J.; Schielen, W.J.G. Assessment of the functional affinity constant of monoclonal antibodies using an improved enzyme-linked immunosorbent assay. *J. Immunol. Methods* **1995**, 184, 207-217.

(5) Underwood, P.A. Problems and pitfalls with measurement of antibody affinity using solid phase binding in the ELISA. *J. Immunol. Methods* **1993**, 164, 119-130.

(6) Pokrić, B.; Pučar, Z. The two-cross immunodiffusion technique for determining diffusion coefficients and precipitating titers of antigen and antibody. *Meth. Enzymol.* **1981**, 73, 306-319.

(7) Živković, T.; Pokrić, B.; Pučar, Z. Determination of precipitating titers and diffusion coefficients by double diffusion in gels. *Anal. Chem.* **1976**, 48, 1408-1412.

(8) Pokrić, B.; Pučar, Z. Microgravimetric determination of precipitable antigens and antibodies in native biological fluids. *J. Immunol. Meth.* **1992**, 148, 49-56.

(9) Juroš, S.; Pučar, Z.; Pokrić, B. Quantitative and qualitative characterization of virus envelope proteins and specific polyclonal antibodies. *J. Biochem. Biophys. Meth.* **1993**, 27, 65-75.

(10) Nielsen, A.E.; Hunding, A.; Pokrić, B. Kinetics of precipitation in gel. *Croat. Chem. Acta* **1997**, 50, 39-64.

(11) Ouchterlony, O; Nilsson, L.-A. Immunodiffusion and immunoelectrophoresis. In: *Handbook of Experimental Immunology*; Weir, D.M., Ed.; J.B. Lippincott Co.: Philadelphia, **1978**, pp. 19.1.-19.44.

(12) Srzić, D.; Pokrić, B.; Pučar, Z. Precipitation in gels under conditions of double diffusion: Critical concentrations and solubility products of salts. *Z. Physik. Chem. (Neue Folge)* **1976**, 103, 157-164.

(13) van Oss, C.J. Specifically impermeable precipitate membranes formed through double diffusion in gels: behaviour with complex forming and with simple system. *J. Colloid Interface Sci.* **1968**, 27, 684-690.

(14) Živković, T.; Pokrić, B.; Pučar, Z. Precipitation under conditions of double diffusion. *J. Chem. Soc. Faraday Trans. I.* **1974**, 70, 1991-1998.

(15) Fuchs, H.; Orberger, G.; Tauber, R.; Gessner, R. Direct calibration ELISA: A rapid method for the simplified determination of association constants of unlabeled biological molecules. *J. Immunol. Meth.* **1995**, 188, 197-208.

(16) Pokrić, B.; Pučar, Z. Dependence of diffusion coefficients and immunoprecipitating titers on pH: Albumin and IgG in human serum and their rabbit antibodies. *Anal. Biochem.* **1980**, 106, 389-396.

(17) Zadro, R.; Pokrić, B.; Pučar, Z. Dependence of diffusion coefficients and immunocrecipitating titers on pH: Human serum transferrin, Immunoglobulin A, human chorionic somatomammotropin, and their rabbit antibodies. *Anal. Biochem.* **1981**, 117, 238-244.

(18) van Oss, C.J. Aspecific and specific intermolecular interactions in aqueous media. *J. Molec. Recognit.* **1990**, 3, 128-136.

(19) van Oss, C.J.; Good, R.J. Orientation of the water molecules of hydration of serum albumin. *J. Prot. Chem.* **1988**, 7, 179-183.

(20) Arend, W.P.; Mannik, M. Determination of soluble immune complex molar composition and antibody association constants by ammonium sulphate precipitation. *J. Immunol.* **1974**, 112, 451-461.

(21) Hudson, B.W. An investigation of the effect of antigen concentration on protein antigen-antiprotein association constants. *Immunochemistry* **1968**, 5, 87-105.

(22) Mattes, M.J. Binding parameters of antibodies reacting with multivalent antigens: Functional affinity or pseudo-affinity. *J. Immunol. Meth.* **1995**, 202, 97-101.

(23) van Oss, C.J.; Absolom, D.R.; Michaeli, I. A disquisition on the energetics of immunoglobulin binding to receptors *in vivo* and *in vitro*. *Immun. Invest.* **1985**, 14, 167-176.

(24) van Oss, C.J.; Walker, J. Concentration dependence of the binding constant of antibodies. *Molec. Immunol.* **1987**, 24, 43-53.

(25) Werblin, T.P.; Siskind, G.W. Distribution of antibody affinities: Technique of measurement. *Immunochemistry*, **1972**, 9, 987-10ll.

(26) Ong, G.L.; Mattes, M.J. Re-evaluation of the concept of functional affinity as applied to bivalent antibody binding to the cell surface antigens. *Mol. Immunol.* **1993**, 30, 1455-1462.

ACKNOWLEDGEMENT

Table 1 The Precipitating Titers of the Antigen (PTA) and Antibody (PTB) Solutions, and the Minimal Dilution Factors (m_0) in antiHSA/HSA Systems.

immune serum	pH	20 °C			40 °C		
		PTA	PTB	m_0	PTA	PTB	m_0
1	5.0	356	265	3.95	549	400	1.66
	5.5	509	262	4.90	565	293	3.22
	7.0	509	348	2.71	562	384	1.79
	8.6	659	412	3.48	735	434	2.37
2	5.0	339	377	5.99	387	448	4.93
	5.5	470	373	7.75	551	448	6.37
	7.0	487	470	5.15	551	554	3.72
	8.6	682	604	5.88	718	638	4.86
3	5.0	270	189	3.72	462	321	1.94
	5.5	382	189	4.76	554	271	3.21
	7.4	382	250	3.02	547	349	1.93
	8.6	508	280	3.10	761	407	1.52

The precipitating titers and dilution factors are dimensionless quantities.

Table 2 The Precipitating Concentrations (c_{pr}) and the Ab/Ag Molar Ratio (n) in antiHSA/HSA System Prepared at Equivalence.

immune serum	pH	20 °C			40 °C			n
		$c_{pr}\, 10^3$ [&]		n [=]	$c_{pr}\, 10^3$ [&]		n [#]	
		A	B		A	B		
1	5.0	1.4045	6.6868	2.06	0.9108	4.4300	2.11	2.085
	5.5	0.9823	6.7633	2.98	0.8850	6.0477	2.96	2.970
	7.0	0.9823	5.0919	2.25	0.8897	4.6145	2. 25	2.250
	8.6	0.7194	4.3009	2.59	0.6803	4.0829	2.60	2.595
2	5.0	1.4749	7.1618	2.10	1.2920	6.0268	2.02	2.060
	5.5	1.0638	7.2386	2.95	0.9074	6.0268	2.88	2.915
	7.0	1.0267	5.7447	2.43	0.9074	4.8736	2.33	2.380
	8.6	0.7331	4.4702	2.64	0.6964	4.2320	2.63	2.635
3	5.0	1.8519	7.1058	1.67	1.0823	4.1838	1.68	1.675
	5.5	1.3089	7.1058	2.35	0.9025	4.9557	2.38	2.365
	7.0	1.3089	5.3720	1.78	0.9141	3.8481	1.82	1.800
	8.6	0.9843	4.7964	2.11	0.6570	3.2998	2.18	2.145

[&] The precipitating concentrations, expressed in g/l, are calculated (eqs 7 and 8) taking into account that the concentration of HSA solutions (A) amounted to $cA_0 = 0.500$ g/l/ and the concentrations of anti HSA in rabbit sera (B) amounted to $cB_0 = 1.772$ g/l (rabbit 1), $cB_0 = 2.700$ g/l, (rabbit 2), and $cB_0 = 1.343$ g/l (rabbit 3).[8]

[=] n is calculated (eq 14) taking into account that molecular mass of antigen and antibodies amounts to 65 000 Da and 16 0000 Da, respectively.

Table 3 The Precipitating Titers of Antigen Solution (PTA*) Referring to the Equilibrium Conditions and the Equilibrium Constants of antiHSA/HSA Reactions (Ka).

immune serum	pH	20 °C		40 °C	
		PTA* [&]	$Ka \ 10^{-4}$ [#]	PTA* [&]	$Ka \ 10^{-4}$ [#]
1	5.0	1406	3.23	911	2.60
2	5.0	2031	3.89	1908	3.76
3	5.0	1004	2.75	869	2.60
1	5.5	2494	4.13	1819	3.53
2	5.5	3643	5.00	3510	4.19
3	5.5	1818	3.62	1778	3.58
1	7.0	1379	3.17	1006	2.71
2	7.0	2508	4.25	2066	3.85
3	7.0	1154	2.95	1056	2.82
1	8.6	2419	4.14	1742	3.51
2	8.6	4010	5.13	3490	4.95
3	8.6	1575	3.41	1157	2.92

[&] The precipitating titers are dimensionless quantities.

[#] The equilibrium constants are expressed in l/mol and calculated according to eqs 5 and 22. For this purpose the precipitating concentrations of antigen solutions in g/l (Table 2) are transformed to equilibrium conditions by using dilution factor m_0 (Table 1) and expressed in mol/l, taking into account that molecular mass of antigen is 65 000 Da. The average of n values determined at 20 and 40 °C (Table 2) is introduced into eqs 5 and 22.
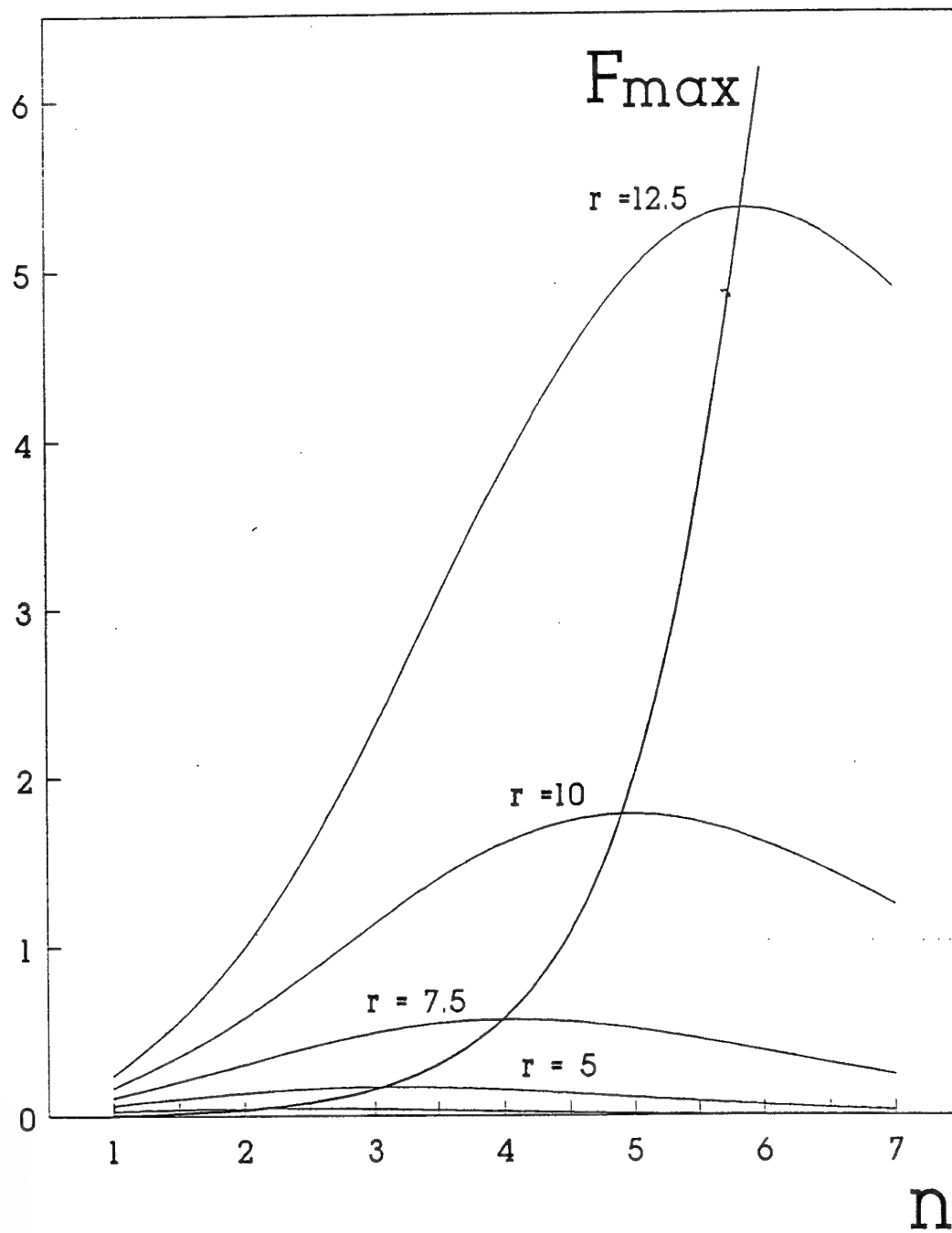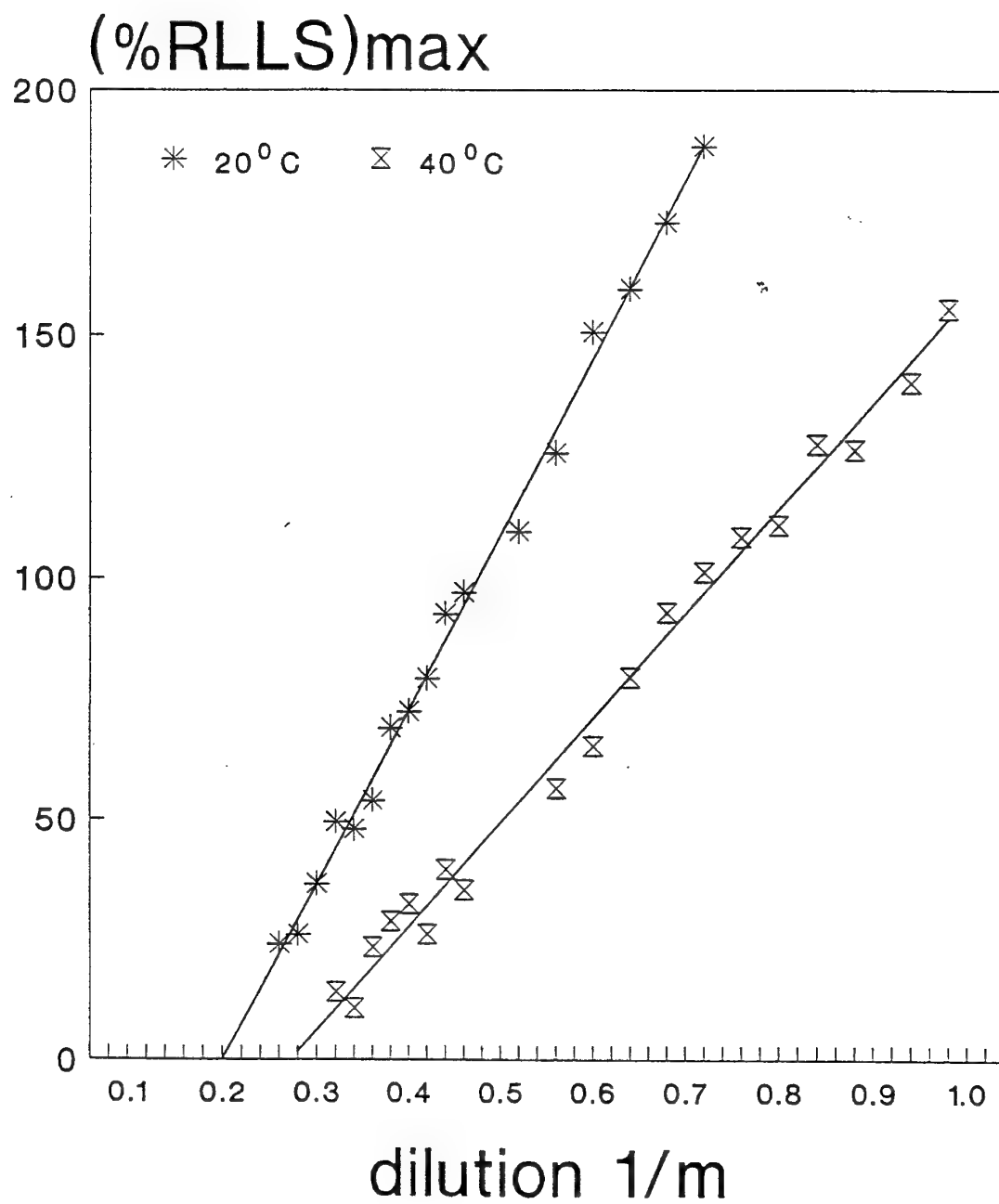
Lists of Face-regular Polyhedra

Gunnar Brinkmann

Fakultät für Mathematik

Universität Bielefeld

D 33501 Bielefeld *

Michel Deza

CNRS and LIENS-DMI

Ecole Normale Supérieure

75005 Paris, France †

0.1 Abstract

We introduce a new notion that connects the combinatorial concept of regularity with the geometrical notion of face-transitivity. This new notion implies finiteness results in case of bounded maximal face size. We give lists of structures for some classes and investigate polyhedra with constant vertex degree and faces of only two sizes.

1 Introduction

A planar (finite or infinite) graph is called *face-transitive*, if the automorphism group acts transitively on the set of faces. For finite polyhedra (see [Ma71]) as well as for infinite graphs in the plane with finite faces and finite vertex degree (that is tilings, see [Ba90][De90]) it is well known that the graph can be realized with its full combinatorial automorphism group as its group of geometrical symmetries. Restricting the attention to polyhedra with constant vertex degree, up to combinatorial equivalence only the 5 *Platonic solids* have an automorphism group acting transitively on their faces. In the remaining text we will restrict our attention to polyhedra with constant vertex degree.

A natural generalisation of this concept – let us call it *weakly face-transitive* – is to require that only faces of the same size are equivalent under the automorphism group. If we define the 0-th corona of a face to be the face itself and the n-th corona to be the set of all those faces that are contained in the (n-1)-th corona or share an edge with it, we can further relax this concept and only require some coronas of fixed size to be isomorphic by an isomorphism mapping the central faces onto each other. A polyhedron with all n-coronas of faces of the same size isomorphic is called *weakly n-transitive*. Obviously, all polyhedra are weakly 0-transitive and if a polyhedron is weakly (n+1)-transitive, it is also weakly n-transitive. So the first interesting case to study is the case of weakly 1-transitive polyhedra. Still relaxing this condition by not requiring the first coronas to be isomorphic, but just to be isomorphic as multisets (that is: every face of a given size i must have the same number of neighbours of size i' for every i'), still gives a very restrictive condition

---

*e-mail: gunnar@mathematik.uni-bielefeld.de

†e-mail: deza@dmi.ens.fr

and as we will see, it already implies finiteness in case the maximal size of a face is bounded. We call this condition (strong) *face-regularity*. So the class of all face-regular polyhedra contains all weakly n-face-transitive polyhedra for any $n \geq 1$, and therefore also the weakly face-transitive or even face-transitive ones.

The same concept can be reached by strengthening the notion of a monochromatic regular dual: Let p_i denote the number of i-gons in a given polyhedron. We use the notation $p = (p_3, p_4, ..., p_i, ...p_b)$ for the face-vector (or p-vector) of a polyhedron; b is the maximal number for which a face with size b exists.

A less restrictive definition of face-regularity, but only for bifaced polyhedra, was considered in [DGr97c]. Namely, if only p_a and p_b are non-zeros and $a < b$, then the number f of i-faces, edge-adjacent to any given i-face, was required to be independent of the choice of the i-face, for i either a or b. For a k-valent polyhedron we write aR_f or bR_f, if this partial (or *weak*) face-regularity holds for a-gonal or, respectively, b-gonal faces. All such simple polyhedra with $b \leq 6$, as well as all 4-valent ones with $b = 4$, except the cases $4R_0$ for $(k; a, b) = (3; 4, 6)$ and aR_0, aR_1 for $(k; a, b) \in \{(4; 3, 4), (3; 5, 6)\}$ were found in [DGr97c]. For example, all 12 (resp. 6,4,10,26) polyhedra bR_f for all five possible cases - $k = 4$; $k = 3, b < 6$ and $k = 3, b = 6, a \in \{3, 4, 5\}$ - are listed there. (The graphs of all 26 $6R_f$ *fullerenes* (i.e. $(k, a, b) = (3, 5, 6)$) are given in list 7 below.) In these cases 8 (resp. 6,4,9,12) polyhedra are also aR_f, i.e. face-regular in the sense of the present paper.

The face-regularity which we consider, is a purely combinatorial property of the skeleton of a polyhedron. It is different from the affine notion of *regular-faced* (i.e. all faces being regular polygons) polyhedra.

We use the abbreviation *frp* for *face-regular polyhedron*. An frp in one of the lists below is described by i_j, where j is the number of the List and i is its number in List j. We also use the notation i for i_1.

We call two frp *fr-isomers*, if they have the same parameters as frp, i.e. v, the p-vector and the numbers $f(a, b)$, i.e. the number of b-faces, edge-adjacent to each a-face for any a, b, coincide.

All fr-isomers in List 1 are bifaced. They are: $11, 12(v = 16)$; $20, 21(v = 32)$; $32, 33(v = 80)$ and 3-faced $49, 50(v = 20)$

All fr-isomers in List 2 are:

for $v = 20$: $10_2, 11_2$;

for $v = 24$: $61_2, 62_2$;

for $v = 26$: 16_2-19_2;

for $v = 28$: $66_2, 67_2$; $69_2, 70_2$; 72_2-74_2; $75_2, 77_2$; $76_2, 78_2$;

for $v = 32$: 28_2-31_2; 32_2-34_2; $87_2, 88_2$; $89_2, 90_2$;

for $v = 36$: $2_2, 3_2$; $95_2, 96_2$; $102_2, 103_2$;

for $v = 40$: $42_2, 43_2$.

for $v = 44$: $5_2, 6_2$; 49_2-51_2; $119_2, 120_2$; 137_2-139_2

All fr-isomers in Lists 4 and 5 are $6_4, 7_4$ with $v = 14$.

Considering the polyhedra of Lists 1,2 and 3 with respect to collapsing of all triangular faces to points, (i.e. the inverse to vertex-truncation), we see that in List 1, any such collapsing gives a member of List 1. But in List 2 there are polyhedra, such that this collapsing does not give an frp. The smallest one is 116_2.

Examples of sequences of frp, such that each of them comes from the previous one by 1-edge truncation are: $1, 4, 2, 6, 7, 8, 9$; $1, 4, 35, 36, 59, 39, 11$; and $1, 4, 2, 6, 14, 4_2$.

2

Some infinite families of 3-valent frp:

Bifaced : $Prism_n$ and $Barrel_n$ (i.e. two n-gons separated by two layers of 5-gons);

3-faced : $Prism_n$, $Barrel_n$, truncated on all 2n vertices of both n-gons;

    $Prism_{2n}$, edge-truncated on n disjoint edges of only one n-gon;

    $Prism_{3n}$, edge-truncated on n edges, separated by at least 2 edges, of only one n-gon;

4-faced : $Prism_n$, (vertex-) truncated on all vertices of only one n-gon;

5-faced : $Barrel_n$, truncated on all vertices of only one n-gon.

In fact, many of the frp in the lists are some partial truncations of $Prism_n$ and $Barrel_n$. For example, there are exactly 10 frp, which are partial truncations of the Cube: There are 1 (resp. 3,1,3,1,1) possibilities for truncations on 1 (resp 2,3,4,6,8) vertices.

Remarks:

(i) Among the chiral polyhedra in the lists are, for example, Nrs 41, 61, 62, 63, 100, 104 in List 2; Nr.9 in List 3; and, especially, Nrs 13,22,34 in List 1 and 9 in List 4 with symmetry T, O, I and O, respectively.

(ii) None of the polyhedra in any of our Lists has a trivial symmetry group.

The Finiteness of Classes with Bounded Face Size

Theorem 1 *For every $n \in \mathbb{N}$ there is only a finite number of face-regular polyhedra with constant vertex degree and face sizes not exceeding n.*

Proof.

We will assume that the polyhedra in question all contain an n-gon. The total number can be obtained by summing up over all $m \leq n$.

Remind that for $i, j \in \mathbb{N}$ the number $f(i, j)$ denotes the number of neighbouring j-gons of an i-gon. So $f(i,j)p_i = f(j,i)p_j$ is the number of edges between i-gons and j-gons and we can express p_j as $p_j = \frac{f(i,j)}{f(j,i)}p_i$ in case i-gonal and j-gonal faces share at least one edge.

Look at the *f-graph* G with vertex set $V = \{i | p_i > 0\}$ and edge set $E = \{\{i,j\} | f(i,j) > 0\}$. This graph is connected since the dual of the underlying polyhedron is connected. We can express every other value p_i by a formula of the kind

$$\frac{f(i_1,i)}{f(i,i_1)}\frac{f(i_2,i_1)}{f(i_1,i_2)} \cdots \frac{f(i_b,i_k)}{f(i_k,i_b)}p_n =: g(i)p_n$$

if $i, i_1, \ldots, i_k, n$ is a (e.g. shortest) path from i to n in G.

Since for fixed n all the $f(i, j)$ as well as the length of the path are bounded and since the number of graphs on n vertices is also finite, we have only a finite number of possible sets of equations $p_i = g(i)p_n$ $(3 \leq i \leq n)$.

As a well known consequence of Euler's formula we get $\sum_{i=3}^{n}(6-i)p_i = 12$ in the 3-valent case, $\sum_{i=3}^{n}(4 - i)p_i = 8$ for 4-valent polyhedra and $\sum_{i=3}^{n}(10 - 3i)p_i = 20$ for 5-valent polyhedra.

Substituting p_i by $g(i)p_n$ in this formula, every set of equations gives exactly one solution for p_n and therefore also for each p_i. So for every set of equations there is a

well determined number of faces and therefore there is a maximum number of faces that is possible.

□

Corollary 1 *If in the cubic case the number of non-hexagons is bounded or in the quartic case the number of non-squares is bounded, then there is only a finite number of face-regular polyhedra.*

Proof.

The fact that the number of faces smaller than 6 (resp. 4) is bounded gives an upper bound on the maximum face size, implying the result by the previous theorem.

□

Statistics

In this section we will give some statistics about the number of face-regular polyhedra compared to the number of all polyhedra for some classes.

vertices	polyhedra	face-regular polyhedra
4	1	1
6	1	1
8	2	2
10	5	4
12	14	7
14	50	5
16	233	15
18	1 249	9
20	7 595	33
22	49 566	11
24	339 722	58
26	2 406 841	29
28	17 490 241	99
30	129 664 753	44
32	977 526 957	194
34	7 475 907 149	25
36	57 896 349 553	318

Table 1: Cubic polyhedra

vertices	polyhedra	face-regular polyhedra
8	1	1
10	1	1
12	2	2
14	5	3
16	12	3
18	34	1
20	130	10
22	525	2
24	2 472	8
26	12 400	5
28	65 619	10
30	357 504	7
32	1 992 985	30
34	11 284 042	1
36	64 719 885	22
38	375 126 827	16
40	2 194 439 398	18

Table 2: Cubic polyhedra without triangles

vertices	polyhedra	face-regular polyhedra
4	1	1
6	1	1
8	2	2
10	5	4
12	10	7
14	15	3
16	30	7
18	44	2
20	77	10
24	184	6
26	267	2
28	420	3
30	595	1
32	883	5
38	2 445	1
44	6 319	1
52	19 345	1
56	32 219	2
60	52 293	1
68	128 343	1
80	425 998	2
140	???	1

Table 3: Cubic polyhedra without faces larger than a hexagon. For all vertex numbers not mentioned, no face-regular polyhedra exist.

2   List 1: all 64 face-regular simple polyhedra with $b \leq 6$.

Among the 64 polyhedra of the List, the first three are regular, then there are 31 bifaced ones: six with $b \leq 5$, four 3_n (for n = 12, 16, 16, 26), nine 4_n (for n = 12, 14, 20, 20, 24, 26, 32, 32, 36) and 12 fullerenes 5_n (which are $F_{24}(D_{6d})$, $F_{28}(T_d)$, $F_{32}(D_{3h})$, $F_{38}(C_{3v})$, $F_{44}(T)$, $F_{52}(T)$, $F_{56}(T_d)$, $F_{60}(I_h)$, $F_{68}(T_d)$, $F_{80}(I_h)$, $F_{80}(D_{5h})$, $F_{140}(I)$). Nrs. 35–57 have three types of faces and last seven polyhedra, Nrs. 58–64, have four types of faces.

Among the 64 polyhedra of the List 1, three are regular ones (Tetrahedron, Cube and Dodecahedron), five are semi-regular (3-, 5-, 6-gonal prisms, truncated octahehedron and truncated Icosahedron) and no one is regular-faced from the list of 92 in [Joh66]. But there are three, which are dual to regular-faced snub disphenoid,

3-augmented 3-gonal prism and gyroelongated square dipyramid (last three have number 84, 51 and 17, respectively, in the list of [Joh66]). Together with three regular ones and 3-, 5-gonal prisms, it gives the duals of all eight convex deltahedra.
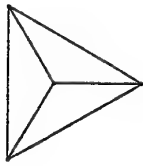
Nr.1 $v = 4$
$p_3 = 4 : 3$
Groupsize: 24
Group: T_d
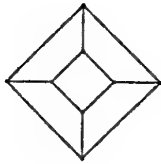
Nr.2 $v = 8$
$p_4 = 6 : 0,4$
Groupsize: 48
Group: O_h
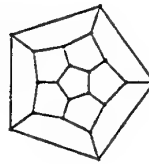
Nr.3 $v = 20$
$p_5 = 12 : 0,0,5$
Groupsize: 120
Group: I_h
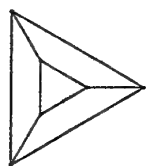
Nr.4 $v = 6$
$p_4 = 3 : 2,2$
$p_3 = 2 : 0,3$
Groupsize: 12
Group: D_{3h}
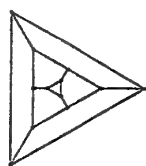
Nr.5 $v = 12$
$p_5 = 6 : 1,0,4$
$p_3 = 2 : 0,0,3$
Groupsize: 12
Group: D_{3d}
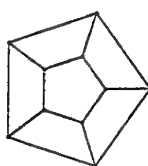
Nr.6 $v = 10$
$p_5 = 2 : 0,5,0$
$p_4 = 5 : 0,2,2$
Groupsize: 20
Group: D_{5h}
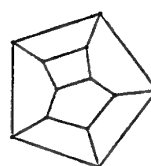
Nr.7 $v = 12$
$p_5 = 4 : 0,3,2$
$p_4 = 4 : 0,1,3$
Groupsize: 8
Group: D_{2d}
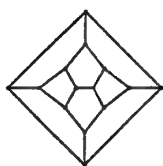
Nr.8 $v = 14$
$p_5 = 6 : 0,2,3$
$p_4 = 3 : 0,0,4$
Groupsize: 12
Group: D_{3h}
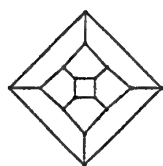
Nr.9 $v = 16$
$p_5 = 8 : 0,1,4$
$p_4 = 2 : 0,0,4$
Groupsize: 16
Group: D_{4d}
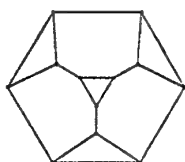
Nr.10 $v = 12$
$p_6 = 4 : 3,0,0,3$
$p_3 = 4 : 0,0,0,3$
Groupsize: 24
Group: T_d
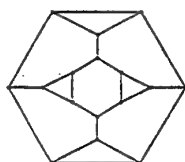
Nr.11 $v = 16$
$p_6 = 6 : 2,0,0,4$
$p_3 = 4 : 0,0,0,3$
Groupsize: 24
Group: T_d
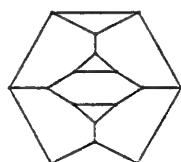
Nr.12 $v = 16$
$p_6 = 6 : 2,0,0,4$
$p_3 = 4 : 0,0,0,3$
Groupsize: 8
Group: D_{2h}
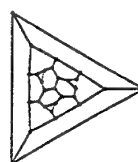
Nr.13 $v = 28$
$p_6 = 12 : 1,0,0,5$
$p_3 = 4 : 0,0,0,3$
Groupsize: 12
Group: T
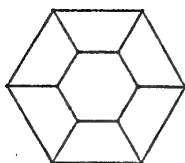
Nr.14 $v = 12$
$p_6 = 2 : 0,6,0,0$
$p_4 = 6 : 0,2,0,2$
Groupsize: 24
Group: D_{6h}
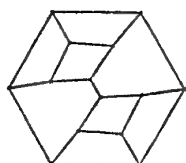
Nr.15 $v = 14$
$p_6 = 3 : 0,4,0,2$
$p_4 = 6 : 0,2,0,2$
Groupsize: 12
Group: D_{3h}
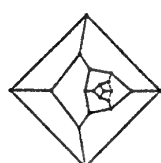
Nr.16 $v = 20$
$p_6 = 6 : 0,2,0,4$
$p_4 = 6 : 0,2,0,2$
Groupsize: 12
Group: D_{3d}
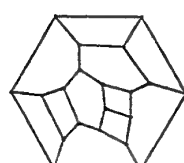
Nr.17 $v = 20$
$p_6 = 6 : 0,3,0,3$
$p_4 = 6 : 0,1,0,3$
Groupsize: 6
Group: S_6
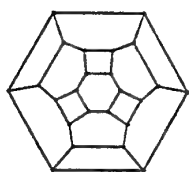
Nr.18 $v = 24$
$p_6 = 8 : 0,3,0,3$
$p_4 = 6 : 0,0,0,4$
Groupsize: 48
Group: O_h
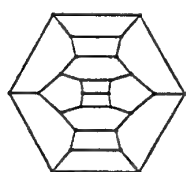
Nr.19 $v = 26$
$p_6 = 9 : 0,2,0,4$
$p_4 = 6 : 0,1,0,3$
Groupsize: 12
Group: D_{3h}
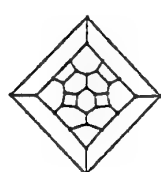
Nr.20 $v = 32$
$p_6 = 12 : 0,2,0,4$
$p_4 = 6 : 0,0,0,4$
Groupsize: 12
Group: D_{3d}
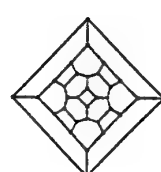
Nr.21 $v = 32$
$p_6 = 12 : 0,2,0,4$
$p_4 = 6 : 0,0,0,4$
Groupsize: 48
Group: O_h
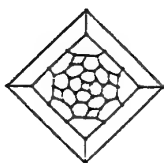
Nr.22 $v = 56$
$p_6 = 24 : 0,1,0,5$
$p_4 = 6 : 0,0,0,4$
Groupsize: 24
Group: O
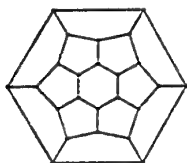
Nr.23 $v = 24$
$p_6 = 2 : 0,0,6,0$
$p_5 = 12 : 0,0,4,1$
Groupsize: 24
Group: D_{6d}
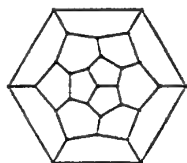
Nr.24 $v = 28$
$p_6 = 4 : 0,0,6,0$
$p_5 = 12 : 0,0,3,2$
Groupsize: 24
Group: T_d
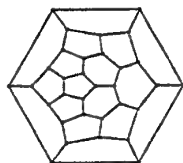
Nr.25 $v = 32$
$p_6 = 6 : 0,0,4,2$
$p_5 = 12 : 0,0,3,2$
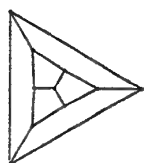Groupsize: 12
Group: D_{3h}
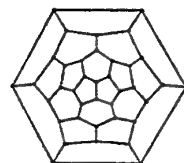
Nr.26 $v = 38$
$p_6 = 9 : 0,0,4,2$
$p_5 = 12 : 0,0,2,3$
Groupsize: 6
Group: C_{3v}
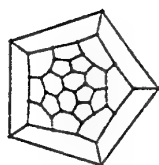
Nr.27 $v = 44$
$p_6 = 12 : 0,0,3,3$
$p_5 = 12 : 0,0,2,3$
Groupsize: 12
Group: T
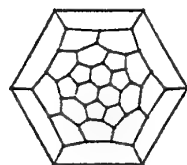
Nr.28 $v = 52$
$p_6 = 16 : 0,0,3,3$
$p_5 = 12 : 0,0,1,4$
Groupsize: 12
Group: T
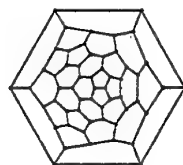
Nr.29 $v = 56$
$p_6 = 18 : 0,0,2,4$
$p_5 = 12 : 0,0,2,3$
Groupsize: 24
Group: T_d
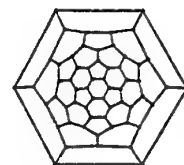
Nr.30 $v = 60$
$p_6 = 20 : 0,0,3,3$
$p_5 = 12 : 0,0,0,5$
Groupsize: 120
Group: I_h
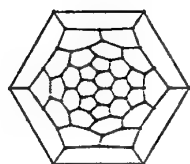
Nr.31 $v = 68$
$p_6 = 24 : 0,0,2,4$
$p_5 = 12 : 0,0,1,4$
Groupsize: 24
Group: T_d
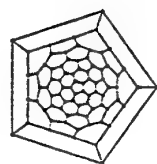
Nr.32 $v = 80$
$p_6 = 30 : 0,0,2,4$
$p_5 = 12 : 0,0,0,5$
Groupsize: 120
Group: I_h
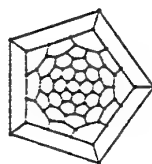
Nr.33 $v = 80$
$p_6 = 30 : 0,0,2,4$
$p_5 = 12 : 0,0,0,5$
Groupsize: 20
Group: D_{5h}
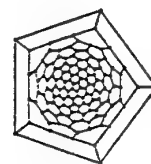
Nr.34 $v = 140$
$p_6 = 60 : 0,0,1,5$
$p_5 = 12 : 0,0,0,5$
Groupsize: 60
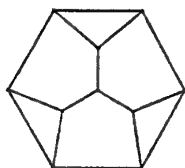Group: I

9
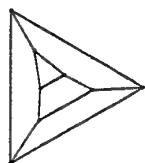
Nr.35 $v = 8$
$p_5 = 2 : 2,2,1$
$p_4 = 2 : 1,1,2$
$p_3 = 2 : 0,1,2$
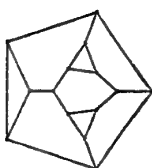Groupsize: 4
Group: C_{2v}

Nr.36 $v = 10$
$p_5 = 3 : 1,2,2$
$p_4 = 3 : 0,2,2$
$p_3 = 1 : 0,0,3$
Groupsize: 6
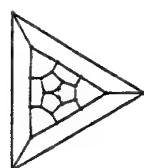Group: C_{3v}
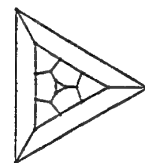
Nr.37 $v = 18$
$p_6 = 6 : 1,2,0,3$
$p_4 = 3 : 0,0,0,4$
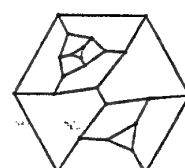$p_3 = 2 : 0,0,0,3$
Groupsize: 12
Group: D_{3h}

Nr.38 $v = 10$
$p_6 = 1 : 3,0,3,0$
$p_5 = 3 : 2,0,2,1$
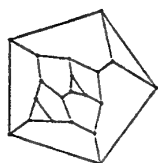$p_3 = 3 : 0,0,2,1$
Groupsize: 6
Group: C_{3v}

Nr.39 $v = 14$
$p_6 = 3 : 2,0,2,2$
$p_5 = 3 : 1,0,2,2$
$p_3 = 3 : 0,0,1,2$
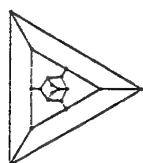Groupsize: 6
Group: C_{3v}

Nr.40 $v = 24$
$p_6 = 6 : 1,0,3,2$
$p_5 = 6 : 0,0,2,3$
$p_3 = 2 : 0,0,0,3$
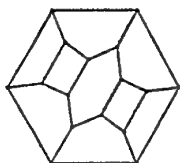Groupsize: 12
Group: D_{3d}

Nr.41 $v = 24$
$p_6 = 6 : 0,0,2,4$
$p_5 = 6 : 1,0,2,2$
$p_3 = 2 : 0,0,3,0$
Groupsize: 12
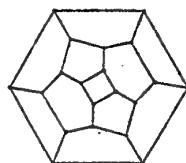Group: D_{3h}

Nr.42 $v = 20$
$p_6 = 6 : 1,0,2,3$
$p_5 = 3 : 1,0,0,4$
$p_3 = 3 : 0,0,1,2$
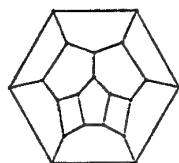Groupsize: 6
Group: C_{3h}

Nr.43 $v = 18$
$p_6 = 3 : 0,0,4,2$
$p_5 = 6 : 1,0,2,2$
$p_3 = 2 : 0,0,3,0$
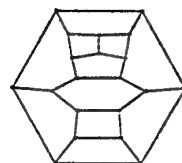Groupsize: 12
Group: D_{3h}

Nr.44 $v = 16$
$p_6 = 2 : 0,2,4,0$
$p_5 = 4 : 0,2,1,2$
$p_4 = 4 : 0,1,2,1$
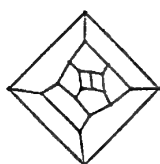Groupsize: 8
Group: D_{2h}

Nr.45 $v = 20$
$p_6 = 3 : 0,2,4,0$
$p_5 = 6 : 0,1,2,2$
$p_4 = 3 : 0,0,2,2$
Groupsize: 12
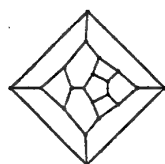Group: D_{3h}

Nr.46 $v = 20$
$p_6 = 4 : 0,2,3,1$
$p_5 = 4 : 0,2,0,3$
$p_4 = 4 : 0,0,2,2$
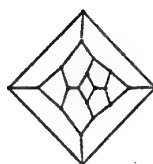Groupsize: 8
Group: D_{2h}

Nr.47 $v = 20$
$p_6 = 4 : 0,1,3,2$
$p_5 = 4 : 0,2,0,3$
$p_4 = 4 : 0,1,2,1$
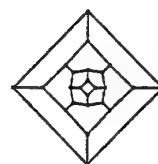Groupsize: 8
Group: D_{2d}

Nr.48 $v = 20$
$p_6 = 4 : 0,2,3,1$
$p_5 = 4 : 0,1,1,3$
$p_4 = 4 : 0,1,1,2$
Groupsize: 4
Group: C_{2v}

Nr.49 $v = 20$
$p_6 = 4 : 0,2,2,2$
$p_5 = 4 : 0,2,1,2$
$p_4 = 4 : 0,0,2,2$
Groupsize: 8
Group: D_{2d}

Nr.50 $v = 20$
$p_6 = 4 : 0,2,2,2$
$p_5 = 4 : 0,2,1,2$
$p_4 = 4 : 0,0,2,2$
Groupsize: 8
Group: D_{2h}

Nr.51 $v = 24$
$p_6 = 4 : 0,0,4,2$
$p_5 = 8 : 0,1,2,2$
$p_4 = 2 : 0,0,4,0$
Groupsize: 16
Group: D_{4h}
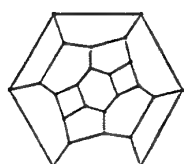
Nr.52 $v = 24$
$p_6 = 6 : 0,2,2,2$
$p_5 = 4 : 0,1,1,3$
$p_4 = 4 : 0,0,1,3$
Groupsize: 4
Group: D_2

Nr.53 $v = 26$
$p_6 = 6 : 0,1,3,2$
$p_5 = 6 : 0,1,1,3$
$p_4 = 3 : 0,0,2,2$
Groupsize: 6
Group: D_3

Nr.54 $v = 28$
$p_6 = 8 : 0,1,1,4$
$p_5 = 4 : 0,2,1,2$
$p_4 = 4 : 0,0,2,2$
Groupsize: 4
Group: D_2

Nr.55 $v = 30$
$p_6 = 10 : 0,2,1,3$
$p_5 = 2 : 0,0,0,5$
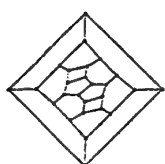$p_4 = 5 : 0,0,0,4$
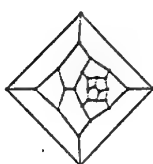Groupsize: 20
Group: D_{5h}

11
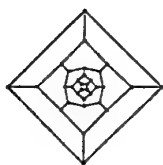
Nr.56　$v = 32$
$p_6 = 8 : 0,0,2,4$
$p_5 = 8 : 0,1,2,2$
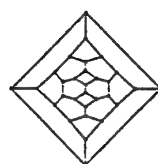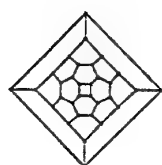$p_4 = 2 : 0,0,4,0$
Groupsize: 16
Group: D_{4h}

Nr.57　$v = 32$
$p_6 = 8 : 0,1,3,2$
$p_5 = 8 : 0,0,2,3$
$p_4 = 2 : 0,0,0,4$
Groupsize: 16
Group: D_{4d}
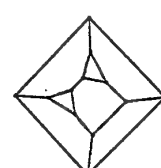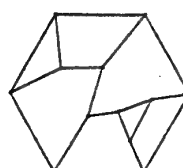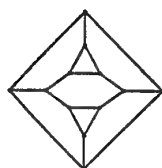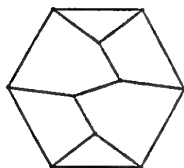
Nr.58　$v = 10$
$p_6 = 1 : 2,2,2,0$
$p_5 = 2 : 1,2,1,1$
$p_4 = 2 : 1,0,2,1$
$p_3 = 2 : 0,1,1,1$
Groupsize: 2
Group: C_2

Nr.59　$v = 12$
$p_6 = 1 : 2,0,4,0$
$p_5 = 4 : 1,1,2,1$
$p_4 = 1 : 0,0,4,0$
$p_3 = 2 : 0,0,2,1$
Groupsize: 4
Group: C_{2v}

Nr.60　$v = 12$
$p_6 = 2 : 1,2,2,1$
$p_5 = 2 : 1,1,1,2$
$p_4 = 2 : 1,0,1,2$
$p_3 = 2 : 0,1,1,1$
Groupsize: 2
Group: C_2

Nr.61　$v = 12$
$p_6 = 2 : 2,1,2,1$
$p_5 = 2 : 1,2,0,2$
$p_4 = 2 : 0,1,2,1$
$p_3 = 2 : 0,0,1,2$
Groupsize: 4
Group: C_{2v}
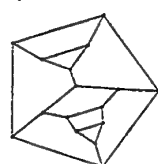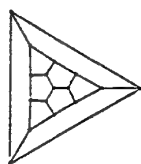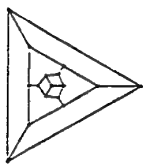
Nr.62　$v = 16$
$p_6 = 3 : 0,2,2,2$
$p_5 = 3 : 1,0,2,2$
$p_4 = 3 : 0,2,0,2$
$p_3 = 1 : 0,0,3,0$
Groupsize: 6
Group: C_{3v}

Nr.63　$v = 16$
$p_6 = 3 : 1,2,1,2$
$p_5 = 3 : 0,2,2,1$
$p_4 = 3 : 0,0,2,2$
$p_3 = 1 : 0,0,0,3$
Groupsize: 6
Group: C_{3v}

Nr.64　$v = 16$
$p_6 = 4 : 1,1,2,2$
$p_5 = 2 : 0,1,0,4$
$p_4 = 2 : 1,0,1,2$
$p_3 = 2 : 0,1,0,2$
Groupsize: 4
Group: C_{2h}

12

3 List 2: all 160 face-regular simple polyhedra with $b = 7$ and up to 24 faces
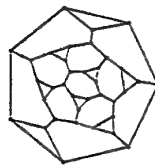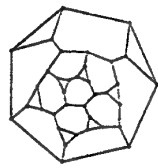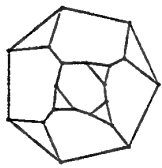
Nr.1 $v = 20$
$p_7 = 6 : 3,0,0,0,4$
$p_3 = 6 : 0,0,0,0,3$
Groupsize: 12

Nr.2 $v = 36$
$p_7 = 12 : 2,0,0,0,5$
$p_3 = 8 : 0,0,0,0,3$
Groupsize: 6

Nr.3 $v = 36$
$p_7 = 12 : 2,0,0,0,5$
$p_3 = 8 : 0,0,0,0,3$
Groupsize: 24
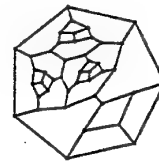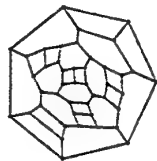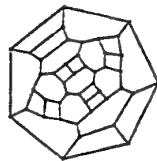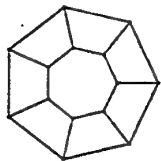
Nr.4 $v = 14$
$p_7 = 2 : 0,7,0,0,0$
$p_4 = 7 : 0,2,0,0,2$
Groupsize: 28

Nr.5 $v = 44$
$p_7 = 12 : 0,3,0,0,4$
$p_4 = 12 : 0,1,0,0,3$
Groupsize: 24

Nr.6 $v = 44$
$p_7 = 12 : 0,3,0,0,4$
$p_4 = 12 : 0,1,0,0,3$
Groupsize: 6

Nr.7 $v = 44$
$p_7 = 12 : 0,2,0,0,5$
$p_4 = 12 : 0,2,0,0,2$
Groupsize: 12
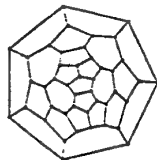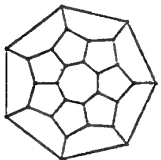
Nr.8 $v = 28$
$p_7 = 2 : 0,0,7,0,0$
$p_3 = 14 : 0,0,4,0,1$
Groupsize: 28

Nr.9 $v = 44$
$p_7 = 6 : 0,0,6,0,1$
$p_5 = 18 : 0,0,3,0,2$
Groupsize: 12

13
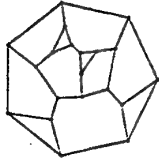
Nr.10 $v = 20$
$p_7 = 4 : 2,0,3,0,2$
$p_5 = 4 : 1,0,1,0,3$
$p_3 = 4 : 0,0,1,0,2$
Groupsize: 4
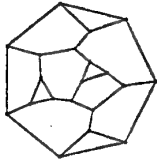
Nr.11 $v = 20$
$p_7 = 4 : 2,0,3,0,2$
$p_5 = 4 : 1,0,1,0,3$
$p_3 = 4 : 0,0,1,0,2$
Groupsize: 4
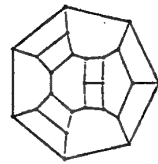
Nr.12 $v = 24$
$p_7 = 4 : 0,4,0,1,2$
$p_6 = 2 : 0,4,0,0,2$
$p_4 = 8 : 0,1,0,1,2$
Groupsize: 8
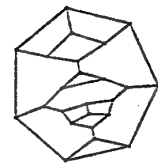
Nr.13 $v = 24$
$p_7 = 6 : 1,2,0,0,4$
$p_4 = 6 : 0,2,0,0,2$
$p_3 = 2 : 0,0,0,0,3$
Groupsize: 4
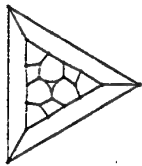
Nr.14 $v = 24$
$p_7 = 6 : 1,3,0,0,3$
$p_4 = 6 : 0,1,0,0,3$
$p_3 = 2 : 0,0,0,0,3$
Groupsize: 6
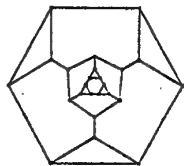
Nr.15 $v = 24$
$p_7 = 6 : 2,0,0,1,4$
$p_6 = 2 : 3,0,0,0,3$
$p_3 = 6 : 0,0,0,1,2$
Groupsize: 12
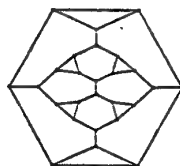
Nr.16 $v = 26$
$p_7 = 6 : 2,0,0,2,3$
$p_6 = 3 : 2,0,0,0,4$
$p_3 = 6 : 0,0,0,1,2$
Groupsize: 4
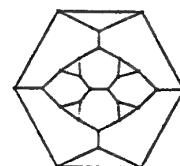
Nr.17 $v = 26$
$p_7 = 6 : 2,0,0,2,3$
$p_6 = 3 : 2,0,0,0,4$
$p_3 = 6 : 0,0,0,1,2$
Groupsize: 12
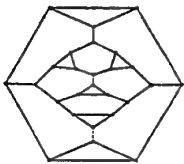
Nr.18 $v = 26$
$p_7 = 6 : 2,0,0,2,3$
$p_6 = 3 : 2,0,0,0,4$
$p_3 = 6 : 0,0,0,1,2$
Groupsize: 4
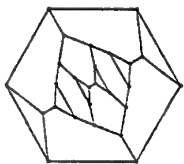
Nr.19 $v = 26$
$p_7 = 6 : 2,0,0,2,3$
$p_6 = 3 : 2,0,0,0,4$
$p_3 = 6 : 0,0,0,1,2$
Groupsize: 12
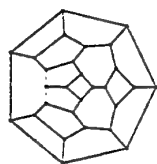
Nr.20 $v = 28$
$p_7 = 4 : 0,2,4,0,1$
$p_5 = 8 : 0,1,2,0,2$
$p_4 = 4 : 0,0,2,0,2$
Groupsize: 8
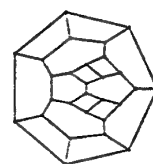
Nr.21 $v = 28$
$p_7 = 4 : 0,4,0,3,0$
$p_6 = 4 : 0,2,0,1,3$
$p_4 = 8 : 0,1,0,1,2$
Groupsize: 8

14
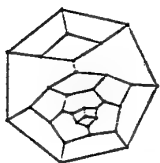
Nr.22 $v = 32$
$p_7 = 6 : 0,2,3,0,2$
$p_5 = 6 : 0,0,2,0,3$
$p_4 = 6 : 0,2,0,0,2$
Groupsize: 12
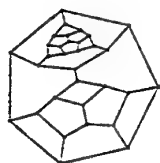
Nr.23 $v = 32$
$p_7 = 6 : 0,2,1,0,4$
$p_5 = 6 : 0,2,2,0,1$
$p_4 = 6 : 0,0,2,0,2$
Groupsize: 12
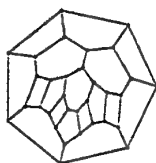
Nr.24 $v = 32$
$p_7 = 6 : 0,2,3,0,2$
$p_5 = 6 : 0,2,0,0,3$
$p_4 = 6 : 0,0,2,0,2$
Groupsize: 12
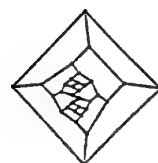
Nr.25 $v = 32$
$p_7 = 6 : 0,1,3,0,3$
$p_5 = 6 : 0,2,0,0,3$
$p_4 = 6 : 0,1,2,0,1$
Groupsize: 6
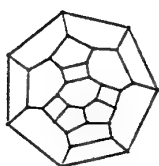
Nr.26 $v = 32$
$p_7 = 6 : 0,3,2,0,2$
$p_5 = 6 : 0,1,2,0,2$
$p_4 = 6 : 0,0,1,0,3$
Groupsize: 12
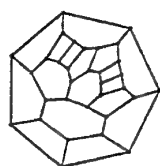
Nr.27 $v = 32$
$p_7 = 6 : 0,2,3,0,2$
$p_5 = 6 : 0,1,1,0,3$
$p_4 = 6 : 0,1,1,0,2$
Groupsize: 6
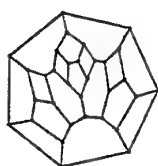
Nr.28 $v = 32$
$p_7 = 6 : 0,2,2,0,3$
$p_5 = 6 : 0,2,1,0,2$
$p_4 = 6 : 0,0,2,0,2$
Groupsize: 12
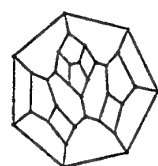
Nr.29 $v = 32$
$p_7 = 6 : 0,2,2,0,3$
$p_5 = 6 : 0,2,1,0,2$
$p_4 = 6 : 0,0,2,0,2$
Groupsize: 4
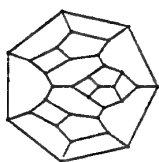
Nr.30 $v = 32$
$p_7 = 6 : 0,2,2,0,3$
$p_5 = 6 : 0,2,1,0,2$
$p_4 = 6 : 0,0,2,0,2$
Groupsize: 12
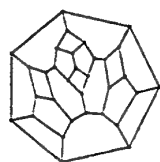
Nr.31 $v = 32$
$p_7 = 6 : 0,2,2,0,3$
$p_5 = 6 : 0,2,1,0,2$
$p_4 = 6 : 0,0,2,0,2$
Groupsize: 4
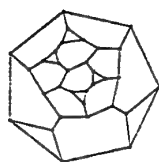
Nr.32 $v = 32$
$p_7 = 6 : 2,0,0,3,2$
$p_6 = 6 : 1,0,0,2,3$
$p_3 = 6 : 0,0,0.1,2$
Groupsize: 6
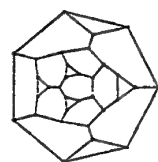
Nr.33 $v = 32$
$p_7 = 6 : 2,0,0,3,2$
$p_6 = 6 : 1,0,0,2,3$
$p_3 = 6 : 0,0,0,1,2$
Groupsize: 12
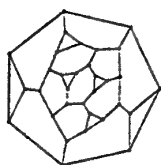
Nr.34 $v = 32$
$p_7 = 6 : 2,0,0,3,2$
$p_6 = 6 : 1,0,0,2,3$
$p_3 = 6 : 0,0,0,1,2$
Groupsize: 6
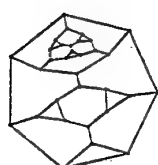
Nr.35 $v = 32$
$p_7 = 6 : 1,0,0,2,4$
$p_6 = 6 : 2,0,0,2,2$
$p_3 = 6 : 0,0,0,2,1$
Groupsize: 12
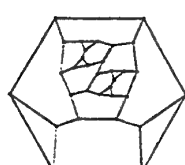
Nr.36 $v = 32$
$p_7 = 6 : 1,0,0,3,3$
$p_6 = 6 : 2,0,0,1,3$
$p_3 = 6 : 0,0,0,2,1$
Groupsize: 6
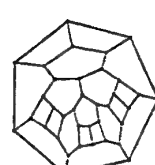
Nr.37 $v = 36$
$p_7 = 4 : 0,2,0,4,1$
$p_6 = 8 : 0,2,0,2,2$
$p_4 = 8 : 0,1,0,2,1$
Groupsize: 8
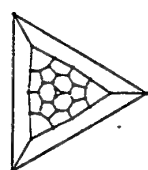
Nr.38 $v = 36$
$p_7 = 6 : 1,0,4,0,2$
$p_5 = 12 : 0,0,3,0,2$
$p_3 = 2 : 0,0,0,0,3$
Groupsize: 12
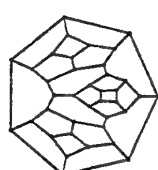
Nr.39 $v = 38$
$p_7 = 6 : 0,0,4,0,3$
$p_5 = 12 : 0,1,2,0,2$
$p_4 = 3 : 0,0,4,0,0$
Groupsize: 12
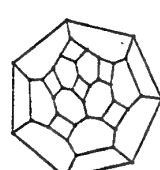
Nr.40 $v = 38$
$p_7 = 6 : 0,3,0,2,2$
$p_6 = 6 : 0,3,0,1,2$
$p_4 = 9 : 0,0,0,2,2$
Groupsize: 12
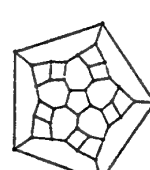
Nr.41 $v = 40$
$p_7 = 10 : 0,3,1,0,3$
$p_5 = 2 : 0,0,0,0,5$
$p_4 = 10 : 0,1,0,0,3$
Groupsize: 10
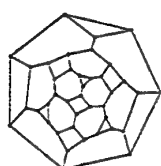
Nr.42 $v = 40$
$p_7 = 12 : 1,2,0,0,4$
$p_4 = 6 : 0,0,0,0,4$
$p_3 = 4 : 0,0,0,0,3$
Groupsize: 6
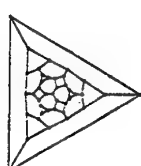
Nr.43 $v = 40$
$p_7 = 12 : 1,2,0,0,4$
$p_4 = 6 : 0,0,0,0,4$
$p_3 = 4 : 0,0,0,0,3$
Groupsize: 24
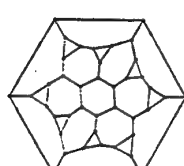
Nr.44 $v = 40$
$p_7 = 12 : 2,0,0,1,4$
$p_6 = 2 : 0,0,0,0,6$
$p_3 = 8 : 0,0,0,0,3$
Groupsize: 8
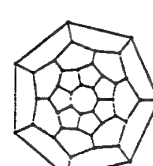
Nr.45 $v = 42$
$p_7 = 2 : 0,0,7,0,0$
$p_6 = 7 : 0,0,4,2,0$
$p_5 = 14 : 0,0,2,2,1$
Groupsize: 28
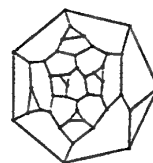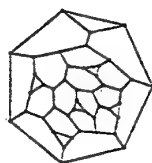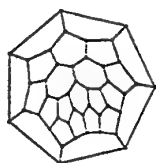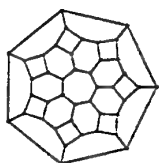
Nr.46 $v = 42$
$p_7 = 2 : 0,0,0,7,0$
$p_6 = 14 : 0,2,0,3,1$
$p_4 = 7 : 0,0,0,4,0$
Groupsize: 28

Nr.47 $v = 44$
$p_7 = 4 : 0,0,4,2,1$
$p_6 = 4 : 0,0,4,0,2$
$p_5 = 16 : 0,0,3,1,1$
Groupsize: 8

Nr.48 $v = 44$
$p_7 = 6 : 1,0,0,4,2$
$p_6 = 12 : 1,0,0,3,2$
$p_3 = 6 : 0,0,0,2,1$
Groupsize: 6

Nr.49 $v = 44$
$p_7 = 12 : 1,0,2,0,4$
$p_5 = 6 : 1,0,0,0,4$
$p_3 = 6 : 0,0,1,0,2$
Groupsize: 6
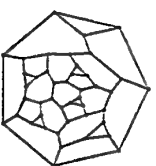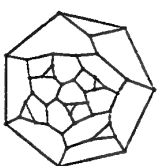
Nr.50 $v = 44$
$p_7 = 12 : 1,0,2,0,4$
$p_5 = 6 : 1,0,0,0,4$
$p_3 = 6 : 0,0,1,0,2$
Groupsize: 6

Nr.51 $v = 44$
$p_7 = 12 : 1,0,2,0,4$
$p_5 = 6 : 1,0,0,0,4$
$p_3 = 6 : 0,0,1,0,2$
Groupsize: 2
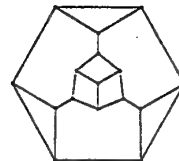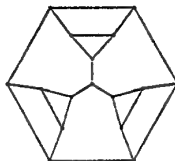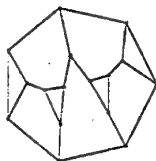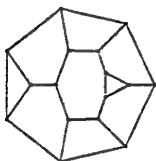
Nr.52 $v = 16$
$p_7 = 2 : 1,2,4,0,0$
$p_5 = 4 : 1,1,1,0,2$
$p_4 = 2 : 0,0,2,0,2$
$p_3 = 2 : 0,0,2,0,1$
Groupsize: 4

Nr.53 $v = 16$
$p_7 = 2 : 2,0,2,2,1$
$p_6 = 2 : 2,0,1,1,2$
$p_5 = 2 : 2,0,0,1,2$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 4

Nr.54 $v = 16$
$p_7 = 3 : 2,2,0,1,2$
$p_6 = 1 : 0,3,0,0,3$
$p_4 = 3 : 1,0,0,1,2$
$p_3 = 3 : 0,1,0,0,2$
Groupsize: 6

Nr.55 $v = 16$
$p_7 = 3 : 2,2,0,1,2$
$p_6 = 1 : 3,0,0,0,3$
$p_4 = 3 : 0,2,0,0,2$
$p_3 = 3 : 0,0,0,1,2$
Groupsize: 6

17
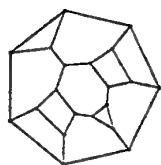
Nr.56 $v = 20$
$p_7 = 2 : 1,2,0,4,0$
$p_6 = 4 : 1,2,0,1,2$
$p_4 = 4 : 0,1,0,2,1$
$p_3 = 2 : 0,0,0,2,1$
Groupsize: 4
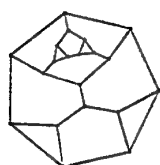
Nr.57 $v = 20$
$p_7 = 2 : 2,0,1,4,0$
$p_6 = 4 : 1,0,1,2,2$
$p_5 = 2 : 2,0,0,2,1$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 4
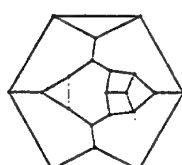
Nr.58 $v = 20$
$p_7 = 3 : 1,2,0,2,2$
$p_6 = 3 : 2,0,0,2,2$
$p_4 = 3 : 0,2,0,0,2$
$p_3 = 3 : 0,0,0,2,1$
Groupsize: 6
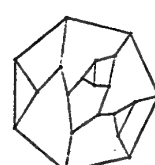
Nr.59 $v = 20$
$p_7 = 3 : 1,2,0,2,2$
$p_6 = 3 : 1,1,0,2,2$
$p_4 = 3 : 1,0,0,1,2$
$p_3 = 3 : 0,1,0,1,1$
Groupsize: 3
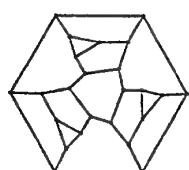
Nr.60 $v = 24$
$p_7 = 3 : 1,0,4,2,0$
$p_6 = 2 : 0,0,3,0,3$
$p_5 = 6 : 1,0,1,1,2$
$p_3 = 3 : 0,0,2,0,1$
Groupsize: 6
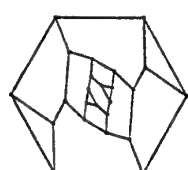
Nr.61 $v = 24$
$p_7 = 4 : 1,0,3,1,2$
$p_6 = 2 : 2,0,2,0,2$
$p_5 = 4 : 1,0,0,1,3$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 4
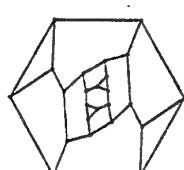
Nr.62 $v = 24$
$p_7 = 4 : 1,0,3,1,2$
$p_6 = 2 : 2,0,2,0,2$
$p_5 = 4 : 1,0,0,1,3$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 4

Nr.63 $v = 24$
$p_7 = 4 : 2,1,0,3,1$
$p_6 = 4 : 1,1,0,1,3$
$p_4 = 2 : 0,0,0,2,2$
$p_3 = 4 : 0,0,0,1,2$
Groupsize: 4
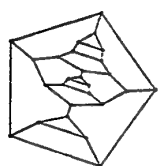
Nr.64 $v = 26$
$p_7 = 6 : 1,1,2,0,3$
$p_5 = 3 : 0,1,0,0,4$
$p_4 = 3 : 1,0,1,0,2$
$p_3 = 3 : 0,1,0,0,2$
Groupsize: 6
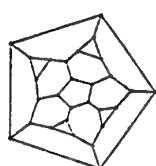
Nr.65 $v = 28$
$p_7 = 2 : 2,0,1,4,0$
$p_6 = 8 : 1,0,1,3,1$
$p_5 = 2 : 0,0,0,4,1$
$p_3 = 4 : 0,0,0,2,1$
Groupsize: 4
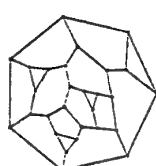
Nr.66 $v = 28$
$p_7 = 4 : 1,0,2,3,1$
$p_6 = 4 : 1,0,2,0,3$
$p_5 = 4 : 1,0,0,2,2$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 4
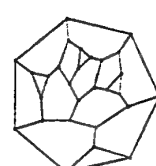
Nr.67 $v = 28$
$p_7 = 4 : 1,0,2,3,1$
$p_6 = 4 : 1,0,2,0,3$
$p_5 = 4 : 1,0,0,2,2$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 4

18
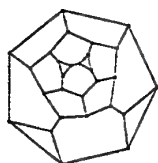
Nr.68 $v = 28$
$p_7 = 4 : 2,0,2,2,1$
$p_6 = 4 : 1,0,2,1,2$
$p_5 = 4 : 0,0,1,2,2$
$p_3 = 4 : 0,0,0,1,2$
Groupsize: 8
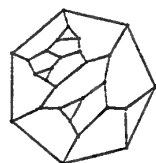
Nr.69 $v = 28$
$p_7 = 4 : 1,0,1,3,2$
$p_6 = 4 : 1,0,2,0,3$
$p_5 = 4 : 1,0,1,2,1$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 4
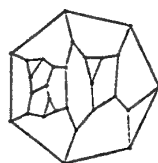
Nr.70 $v = 28$
$p_7 = 4 : 1,0,1,3,2$
$p_6 = 4 : 1,0,2,0,3$
$p_5 = 4 : 1,0,1,2,1$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 4
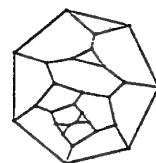
Nr.71 $v = 28$
$p_7 = 4 : 1,0,3,2,1$
$p_6 = 4 : 2,0,1,1,2$
$p_5 = 4 : 0,0,1,1,3$
$p_3 = 4 : 0,0,0,2,1$
Groupsize: 4
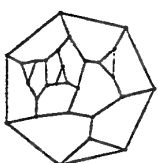
Nr.72 $v = 28$
$p_7 = 4 : 1,0,2,3,1$
$p_6 = 4 : 1,0,1,1,3$
$p_5 = 4 : 1,0,1,1,2$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 2
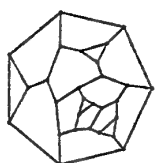
Nr.73 $v = 28$
$p_7 = 4 : 1,0,2,3,1$
$p_6 = 4 : 1,0,1,1,3$
$p_5 = 4 : 1,0,1,1,2$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 4
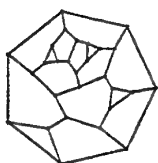
Nr.74 $v = 28$
$p_7 = 4 : 1,0,2,3,1$
$p_6 = 4 : 1,0,1,1,3$
$p_5 = 4 : 1,0,1,1,2$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 4
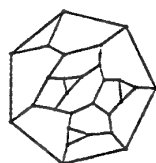
Nr.75 $v = 28$
$p_7 = 4 : 1,0,2,2,2$
$p_6 = 4 : 1,0,2,1,2$
$p_5 = 4 : 1,0,0,2,2$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 4
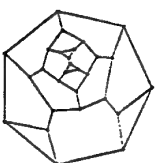
Nr.76 $v = 28$
$p_7 = 4 : 2,0,2,2,1$
$p_6 = 4 : 0,0,2,2,2$
$p_5 = 4 : 1,0,0,2,2$
$p_3 = 4 : 0,0,1,0,2$
Groupsize: 8
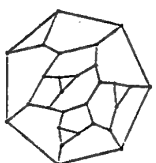
Nr.77 $v = 28$
$p_7 = 4 : 1,0,2,2,2$
$p_6 = 4 : 1,0,2,1,2$
$p_5 = 4 : 1,0,0,2,2$
$p_3 = 4 : 0,0,1,1,1$
Groupsize: 4
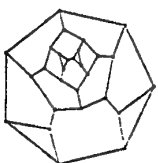
Nr.78 $v = 28$
$p_7 = 4 : 2,0,2,2,1$
$p_6 = 4 : 0,0,2,2,2$
$p_5 = 4 : 1,0,0,2,2$
$p_3 = 4 : 0,0,1,0,2$
Groupsize: 8
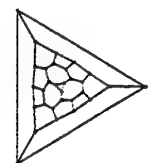
Nr.79 $v = 30$
$p_7 = 6 : 1,1,3,0,2$
$p_5 = 6 : 0,1,1,0,3$
$p_4 = 3 : 0,0,2,0,2$
$p_3 = 2 : 0,0,0,0,3$
Groupsize: 6

19
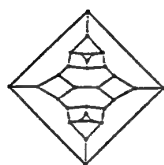
Nr.80 $v = 30$
$p_7 = 6 : 0,2,2,0,3$
$p_5 = 6 : 1,0,2,0,2$
$p_4 = 3 : 0,0,0,0,4$
$p_3 = 2 : 0,0,3,0,0$
Groupsize: 12
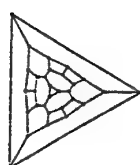
Nr.81 $v = 30$
$p_7 = 6 : 1,2,0,2,2$
$p_6 = 3 : 0,2,0,0,4$
$p_4 = 6 : 0,1,0,1,2$
$p_3 = 2 : 0,0,0,0,3$
Groupsize: 12
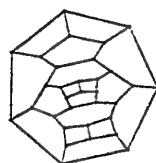
Nr.82 $v = 32$
$p_7 = 3 : 0,2,1,4,0$
$p_6 = 6 : 0,1,1,2,2$
$p_5 = 3 : 0,2,0,2,1$
$p_4 = 6 : 0,1,1,1,1$
Groupsize: 6
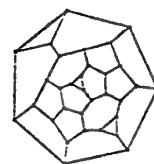
Nr.83 $v = 32$
$p_7 = 3 : 1,0,2,4,0$
$p_6 = 6 : 1,0,2,1,2$
$p_5 = 6 : 0,0,2,2,1$
$p_3 = 3 : 0,0,0,2,1$
Groupsize: 6
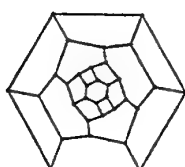
Nr.84 $v = 32$
$p_7 = 4 : 0,1,4,0,2$
$p_6 = 2 : 0,2,4,0,0$
$p_5 = 8 : 0,1,1,1,2$
$p_4 = 4 : 0,0,2,1,1$
Groupsize: 8
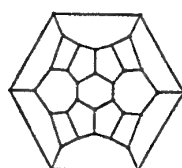
Nr.85 $v = 32$
$p_7 = 4 : 0,2,4,1,0$
$p_6 = 2 : 0,0,4,0,2$
$p_5 = 8 : 0,1,1,1,2$
$p_4 = 4 : 0,0,2,0,2$
Groupsize: 8
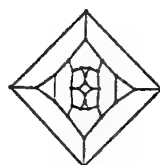
Nr.86 $v = 32$
$p_7 = 4 : 1,0,0,4,2$
$p_6 = 8 : 1,1,0,2,2$
$p_4 = 2 : 0,0,0,4,0$
$p_3 = 4 : 0,0,0,2,1$
Groupsize: 8
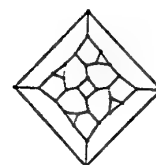
Nr.87 $v = 32$
$p_7 = 8 : 1,1,2,0,3$
$p_5 = 4 : 1,0,0,0,4$
$p_4 = 2 : 0,0,0,0,4$
$p_3 = 4 : 0,0,1,0,2$
Groupsize: 8
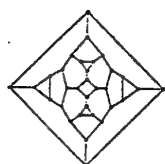
Nr.88 $v = 32$
$p_7 = 8 : 1,1,2,0,3$
$p_5 = 4 : 1,0,0,0,4$
$p_4 = 2 : 0,0,0,0,4$
$p_3 = 4 : 0,0,1,0,2$
Groupsize: 8
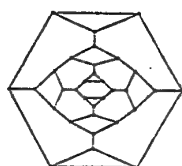
Nr.89 $v = 32$
$p_7 = 8 : 1,2,0,1,3$
$p_6 = 2 : 2,0,0,0,4$
$p_4 = 4 : 0,0,0,0,4$
$p_3 = 4 : 0,0,0,1,2$
Groupsize: 8
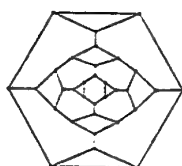
Nr.90 $v = 32$
$p_7 = 8 : 1,2,0,1,3$
$p_6 = 2 : 2,0,0,0,4$
$p_4 = 4 : 0,0,0,0,4$
$p_3 = 4 : 0,0,0,1,2$
Groupsize: 8
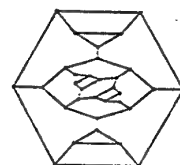
Nr.91 $v = 32$
$p_7 = 8 : 1,1,0,1,4$
$p_6 = 2 : 0,2,0,0,4$
$p_4 = 4 : 1,0,0,1,2$
$p_3 = 4 : 0,1,0,0,2$
Groupsize: 4

20
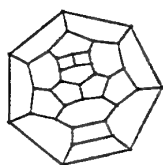
Nr.92 $v = 36$
$p_7 = 4 : 0,1,4,2,0$
$p_6 = 4 : 0,2,2,0,2$
$p_5 = 8 : 0,0,2,1,2$
$p_4 = 4 : 0,1,0,2,1$
Groupsize: 8
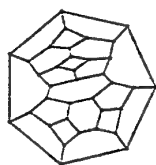
Nr.93 $v = 36$
$p_7 = 4 : 0,0,4,1,2$
$p_6 = 4 : 0,2,2,1,1$
$p_5 = 8 : 0,1,1,1,2$
$p_4 = 4 : 0,0,2,2,0$
Groupsize: 8
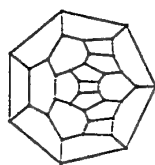
Nr.94 $v = 36$
$p_7 = 4 : 0,1,4,2,0$
$p_6 = 4 : 0,1,2,1,2$
$p_5 = 8 : 0,1,1,1,2$
$p_4 = 4 : 0,0,2,1,1$
Groupsize: 8
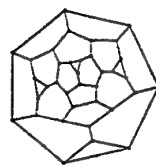
Nr.95 $v = 36$
$p_7 = 4 : 1,0,2,4,0$
$p_6 = 8 : 1,0,1,2,2$
$p_5 = 4 : 0,0,1,2,2$
$p_3 = 4 : 0,0,0,2,1$
Groupsize: 8
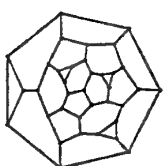
Nr.96 $v = 36$
$p_7 = 4 : 1,0,2,4,0$
$p_6 = 8 : 1,0,1,2,2$
$p_5 = 4 : 0,0,1,2,2$
$p_3 = 4 : 0,0,0,2,1$
Groupsize: 4

Nr.97 $v = 36$
$p_7 = 4 : 0,0,2,4,1$
$p_6 = 8 : 1,0,1,2,2$
$p_5 = 4 : 1,0,0,2,2$
$p_3 = 4 : 0,0,1,2,0$
Groupsize: 8
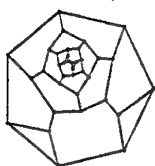
Nr.98 $v = 36$
$p_7 = 4 : 2,0,2,2,1$
$p_6 = 8 : 0,0,1,4,1$
$p_5 = 4 : 1,0,0,2,2$
$p_3 = 4 : 0,0,1,0,2$
Groupsize: 4

Nr.99 $v = 36$
$p_7 = 6 : 0,1,2,0,4$
$p_6 = 2 : 0,3,3,0,0$
$p_5 = 6 : 0,2,0,1,2$
$p_4 = 6 : 0,0,2,1,1$
Groupsize: 12
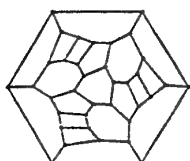
Nr.100 $v = 36$
$p_7 = 6 : 0,2,3,1,1$
$p_6 = 2 : 0,0,3,0,3$
$p_5 = 6 : 0,1,0,1,3$
$p_4 = 6 : 0,1,1,0,2$
Groupsize: 6
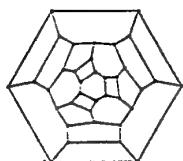
Nr.101 $v = 36$
$p_7 = 6 : 0,2,3,1,1$
$p_6 = 2 : 0,3,0,0,3$
$p_5 = 6 : 0,1,1,0,3$
$p_4 = 6 : 0,0,1,1,2$
Groupsize: 6
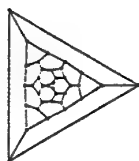
Nr.102 $v = 36$
$p_7 = 6 : 1,2,0,2,2$
$p_6 = 6 : 0,2,0,2,2$
$p_4 = 6 : 0,0,0,2,2$
$p_3 = 2 : 0,0,0,0,3$
Groupsize: 12
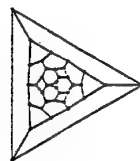
Nr.103 $v = 36$
$p_7 = 6 : 1,2,0,2,2$
$p_6 = 6 : 0,2,0,2,2$
$p_4 = 6 : 0,0,0,2,2$
$p_3 = 2 : 0,0,0,0,3$
Groupsize: 12
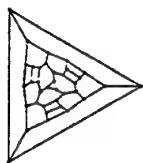
Nr.104 $v = 36$
$p_7 = 6 : 1,1,0,3,2$
$p_6 = 6 : 0,2,0,1,3$
$p_4 = 6 : 0,1,0,2,1$
$p_3 = 2 : 0,0,0,0,3$
Groupsize: 6

Nr.105 $v = 36$
$p_7 = 6 : 1,0,0,2,4$
$p_6 = 6 : 0,2,0,2,2$
$p_4 = 6 : 0,2,0,2,0$
$p_3 = 2 : 0,0,0,0,3$
Groupsize: 4
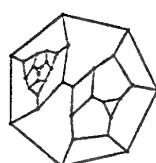
Nr.106 $v = 36$
$p_7 = 6 : 0,2,0,1,4$
$p_6 = 6 : 1,2,0,2,1$
$p_4 = 6 : 0,0,0,2,2$
$p_3 = 2 : 0,0,0,3,0$
Groupsize: 12
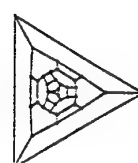
Nr.107 $v = 36$
$p_7 = 6 : 0,3,0,2,2$
$p_6 = 6 : 1,1,0,2,2$
$p_4 = 6 : 0,0,0,1,3$
$p_3 = 2 : 0,0,0,3,0$
Groupsize: 12
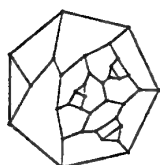
Nr.108 $v = 36$
$p_7 = 8 : 1,1,0,2,3$
$p_6 = 4 : 0,1,0,1,4$
$p_4 = 4 : 1,0,0,1,2$
$p_3 = 4 : 0,1,0,0,2$
Groupsize: 4
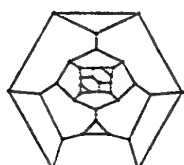
Nr.109 $v = 36$
$p_7 = 8 : 1,1,0,1,4$
$p_6 = 4 : 1,2,0,1,2$
$p_4 = 4 : 0,0,0,2,2$
$p_3 = 4 : 0,0,0,1,2$
Groupsize: 4
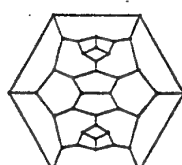
Nr.110 $v = 38$
$p_7 = 6 : 0,2,2,1,2$
$p_6 = 3 : 0,0,4,0,2$
$p_5 = 6 : 0,0,1,2,2$
$p_4 = 6 : 0,2,0,0,2$
Groupsize: 12
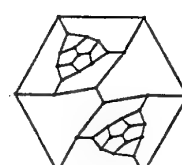
Nr.111 $v = 38$
$p_7 = 6 : 0,2,1,2,2$
$p_6 = 3 : 0,0,0,2,4$
$p_5 = 6 : 0,2,2,0,1$
$p_4 = 6 : 0,0,2,0,2$
Groupsize: 12
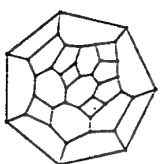
Nr.112 $v = 38$
$p_7 = 6 : 0,2,2,1,2$
$p_6 = 3 : 0,2,2,0,2$
$p_5 = 6 : 0,1,1,1,2$
$p_4 = 6 : 0,0,1,1,2$
Groupsize: 6
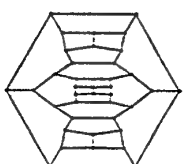
Nr.113 $v = 38$
$p_7 = 6 : 0,1,2,2,2$
$p_6 = 3 : 0,0,2,0,4$
$p_5 = 6 : 0,2,0,1,2$
$p_4 = 6 : 0,1,2,0,1$
Groupsize: 12
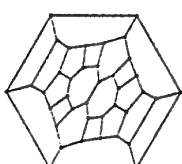
Nr.114 $v = 38$
$p_7 = 6 : 0,2,2,2,1$
$p_6 = 3 : 0,2,0,0,4$
$p_5 = 6 : 0,1,2,0,2$
$p_4 = 6 : 0,0,1,1,2$
Groupsize: 12
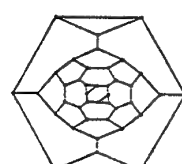
Nr.115 $v = 40$
$p_7 = 8 : 1,0,3,1,2$
$p_6 = 2 : 2,0,0,0,4$
$p_5 = 8 : 0,0,2,0,3$
$p_3 = 4 : 0,0,0,1,2$
Groupsize: 4

22
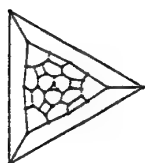
Nr.116 $v = 42$
$p_7 = 6 : 1,1,0,3,2$
$p_6 = 9 : 0,2,0,2,2$
$p_4 = 6 : 0,0,0,3,1$
$p_3 = 2 : 0,0,0,0,3$
Groupsize: 6
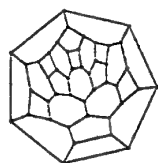
Nr.117 $v = 44$
$p_7 = 6 : 0,1,2,2,2$
$p_6 = 6 : 0,2,2,0,2$
$p_5 = 6 : 0,1,0,2,2$
$p_4 = 6 : 0,0,1,2,1$
Groupsize: 12
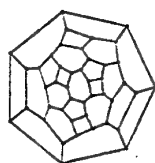
Nr.118 $v = 44$
$p_7 = 6 : 0,2,2,1,2$
$p_6 = 6 : 0,1,2,2,1$
$p_5 = 6 : 0,1,0,2,2$
$p_4 = 6 : 0,0,1,1,2$
Groupsize: 6
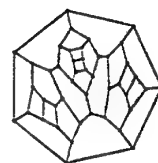
Nr.119 $v = 44$
$p_7 = 6 : 0,0,2,2,3$
$p_6 = 6 : 0,2,2,0,2$
$p_5 = 6 : 0,1,0,2,2$
$p_4 = 6 : 0,1,1,2,0$
Groupsize: 4
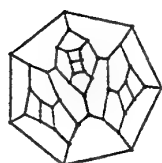
Nr.120 $v = 44$
$p_7 = 6 : 0,0,2,2,3$
$p_6 = 6 : 0,2,2,0,2$
$p_5 = 6 : 0,1,0,2,2$
$p_4 = 6 : 0,1,1,2,0$
Groupsize: 12
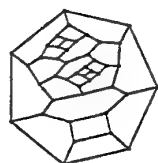
Nr.121 $v = 44$
$p_7 = 6 : 0,1,2,3,1$
$p_6 = 6 : 0,0,1,2,3$
$p_5 = 6 : 0,2,0,1,2$
$p_4 = 6 : 0,1,2,0,1$
Groupsize: 6
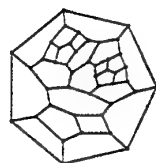
Nr.122 $v = 44$
$p_7 = 6 : 0,1,1,3,2$
$p_6 = 6 : 0,1,1,1,3$
$p_5 = 6 : 0,2,1,1,1$
$p_4 = 6 : 0,0,2,1,1$
Groupsize: 6
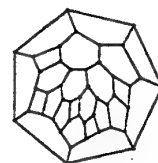
Nr.123 $v = 44$
$p_7 = 6 : 0,1,2,2,2$
$p_6 = 6 : 0,2,1,1,2$
$p_5 = 6 : 0,1,1,1,2$
$p_4 = 6 : 0,0,1,2,1$
Groupsize: 6
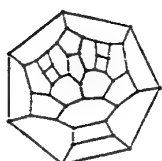
Nr.124 $v = 44$
$p_7 = 6 : 0,1,2,2,2$
$p_6 = 6 : 0,2,2,0,2$
$p_5 = 6 : 0,0,1,2,2$
$p_4 = 6 : 0,1,0,2,1$
Groupsize: 12
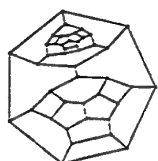
Nr.125 $v = 44$
$p_7 = 6 : 0,1,0,2,4$
$p_6 = 6 : 0,2,2,0,2$
$p_5 = 6 : 0,1,2,2,0$
$p_4 = 6 : 0,0,1,2,1$
Groupsize: 12
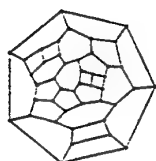
Nr.126 $v = 44$
$p_7 = 6 : 0,1,2,3,1$
$p_6 = 6 : 0,2,1,0,3$
$p_5 = 6 : 0,0,2,1,2$
$p_4 = 6 : 0,1,0,2,1$
Groupsize: 6
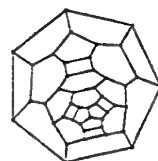
Nr.127 $v = 44$
$p_7 = 6 : 0,2,2,3,0$
$p_6 = 6 : 0,1,0,2,3$
$p_5 = 6 : 0,1,2,0,2$
$p_4 = 6 : 0,0,1,1,2$
Groupsize: 12
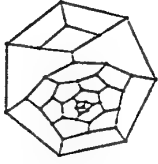
Nr.128 $v = 44$
$p_7 = 6 : 0,2,1,2,2$
$p_6 = 6 : 0,0,3,1,2$
$p_5 = 6 : 0,0,1,3,1$
$p_4 = 6 : 0,2,0,0,2$
Groupsize: 6
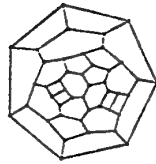
Nr.129 $v = 44$
$p_7 = 6 : 0,2,2,3,0$
$p_6 = 6 : 0,1,1,1,3$
$p_5 = 6 : 0,0,2,1,2$
$p_4 = 6 : 0,1,0,1,2$
Groupsize: 6
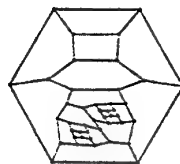
Nr.130 $v = 44$
$p_7 = 6 : 0,0,1,3,3$
$p_6 = 6 : 0,1,2,0,3$
$p_5 = 6 : 0,2,0,2,1$
$p_4 = 6 : 0,1,2,1,0$
Groupsize: 6
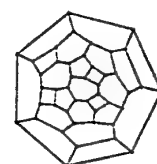
Nr.131 $v = 44$
$p_7 = 6 : 0,2,2,1,2$
$p_6 = 6 : 0,1,2,2,1$
$p_5 = 6 : 0,1,0,2,2$
$p_4 = 6 : 0,0,1,1,2$
Groupsize: 6
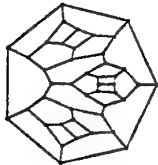
Nr.132 $v = 44$
$p_7 = 6 : 0,0,2,2,3$
$p_6 = 6 : 0,2,2,0,2$
$p_5 = 6 : 0,1,0,2,2$
$p_4 = 6 : 0,1,1,2,0$
Groupsize: 12
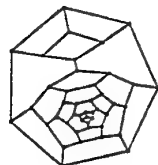
Nr.133 $v = 44$
$p_7 = 6 : 0,2,2,1,2$
$p_6 = 6 : 0,0,3,2,1$
$p_5 = 6 : 0,0,0,3,2$
$p_4 = 6 : 0,2,0,0,2$
Groupsize: 12
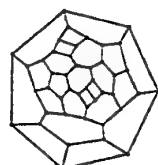
Nr.134 $v = 44$
$p_7 = 6 : 0,2,1,3,1$
$p_6 = 6 : 0,1,2,0,3$
$p_5 = 6 : 0,0,2,2,1$
$p_4 = 6 : 0,1,0,1,2$
Groupsize: 6
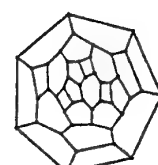
Nr.135 $v = 44$
$p_7 = 6 : 0,2,2,2,1$
$p_6 = 6 : 0,1,1,2,2$
$p_5 = 6 : 0,1,1,1,2$
$p_4 = 6 : 0,0,1,1,2$
Groupsize: 6
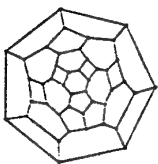
Nr.136 $v = 44$
$p_7 = 6 : 0,2,2,2,1$
$p_6 = 6 : 0,2,1,1,2$
$p_5 = 6 : 0,0,2,1,2$
$p_4 = 6 : 0,0,0,2,2$
Groupsize: 12

Nr.137 $v = 44$
$p_7 = 6 : 0,1,1,3,2$
$p_6 = 6 : 0,1,1,1,3$
$p_5 = 6 : 0,2,1,1,1$
$p_4 = 6 : 0,0,2,1,1$
Groupsize: 6

Nr.138 $v = 44$
$p_7 = 6 : 0,1,1,3,2$
$p_6 = 6 : 0,1,1,1,3$
$p_5 = 6 : 0,2,1,1,1$
$p_4 = 6 : 0,0,2,1,1$
Groupsize: 2

Nr.139 $v = 44$
$p_7 = 6 : 0,1,1,3,2$
$p_6 = 6 : 0,1,1,1,3$
$p_5 = 6 : 0,2,1,1,1$
$p_4 = 6 : 0,0,2,1,1$
Groupsize: 2

24
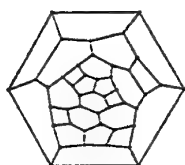
Nr.140 $v = 44$
$p_7 = 6 : 0,2,1,2,2$
$p_6 = 6 : 0,1,2,1,2$
$p_5 = 6 : 0,1,1,2,1$
$p_4 = 6 : 0,0,1,1,2$
Groupsize: 6
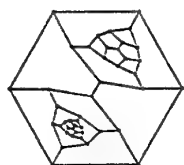
Nr.141 $v = 44$
$p_7 = 6 : 0,2,1,2,2$
$p_6 = 6 : 0,0,0,4,2$
$p_5 = 6 : 0,2,2,0,1$
$p_4 = 6 : 0,0,2,0,2$
Groupsize: 12
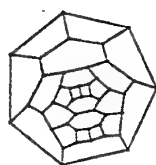
Nr.142 $v = 44$
$p_7 = 6 : 0,2,2,2,1$
$p_6 = 6 : 0,0,2,2,2$
$p_5 = 6 : 0,1,0,2,2$
$p_4 = 6 : 0,1,1,0,2$
Groupsize: 12
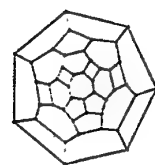
Nr.143 $v = 44$
$p_7 = 6 : 0,2,2,3,0$
$p_6 = 6 : 0,2,1,0,3$
$p_5 = 6 : 0,0,2,1,2$
$p_4 = 6 : 0,0,0,2,2$
Groupsize: 12

Nr.144 $v = 44$
$p_7 = 6 : 0,1,3,1,2$
$p_6 = 6 : 0,2,1,2,1$
$p_5 = 6 : 0,1,0,1,3$
$p_4 = 6 : 0,0,1,2,1$
Groupsize: 6
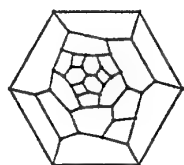
Nr.145 $v = 44$
$p_7 = 6 : 0,1,3,1,2$
$p_6 = 6 : 0,2,1,2,1$
$p_5 = 6 : 0,1,0,1,3$
$p_4 = 6 : 0,0,1,2,1$
Groupsize: 6
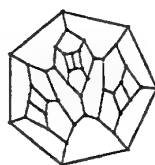
Nr.146 $v = 44$
$p_7 = 6 : 0,0,2,2,3$
$p_6 = 6 : 0,2,2,0,2$
$p_5 = 6 : 0,1,0,2,2$
$p_4 = 6 : 0,1,1,2,0$
Groupsize: 4

Nr.147 $v = 44$
$p_7 = 6 : 0,0,3,2,2$
$p_6 = 6 : 0,2,0,2,2$
$p_5 = 6 : 0,0,2,0,3$
$p_4 = 6 : 0,2,0,2,0$
Groupsize: 12
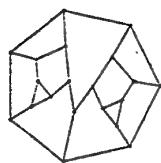
Nr.148 $v = 20$
$p_7 = 2 : 0,2,2,2,1$
$p_6 = 2 : 1,0,2,1,2$
$p_5 = 4 : 1,1,1,1,1$
$p_4 = 2 : 0,0,2,0,2$
$p_3 = 2 : 0,0,2,1,0$
Groupsize: 4
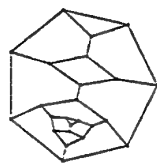
Nr.149 $v = 24$
$p_7 = 2 : 1,0,2,4,0$
$p_6 = 4 : 0,1,1,2,2$
$p_5 = 4 : 1,1,1,1,1$
$p_4 = 2 : 0,0,2,2,0$
$p_3 = 2 : 0,0,2,0,1$
Groupsize: 4
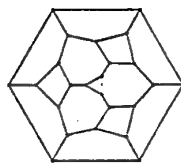
Nr.150 $v = 24$
$p_7 = 3 : 1,2,2,0,2$
$p_6 = 1 : 0,0,6,0,0$
$p_5 = 6 : 0,1,2,1,1$
$p_4 = 3 : 0,0,2,0,2$
$p_3 = 1 : 0,0,0,0,3$
Groupsize: 6
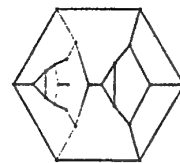
Nr.151 $v = 24$
$p_7 = 4 : 1,1,1,2,2$
$p_6 = 2 : 0,2,0,0,4$
$p_5 = 2 : 1,2,0,0,2$
$p_4 = 4 : 0,1,1,1,1$
$p_3 = 2 : 0,0,1,0,2$
Groupsize: 4
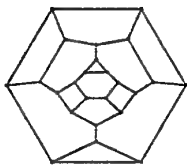
Nr.152 $v = 24$
$p_7 = 4 : 1,2,1,1,2$
$p_6 = 2 : 1,2,1,0,2$
$p_5 = 2 : 0,2,0,1,2$
$p_4 = 4 : 0,0,1,1,2$
$p_3 = 2 : 0,0,0,1,2$
  Groupsize: 4
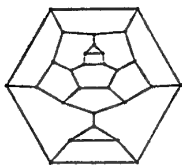
Nr.153 $v = 28$
$p_7 = 4 : 1,1,3,1,1$
$p_6 = 2 : 0,1,3,0,2$
$p_5 = 6 : 0,0,2,1,2$
$p_4 = 2 : 1,0,0,1,2$
$p_3 = 2 : 0,1,0,0,2$
  Groupsize: 4
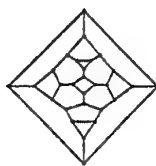
Nr.154 $v = 28$
$p_7 = 4 : 1,2,1,2,1$
$p_6 = 4 : 0,2,1,1,2$
$p_5 = 2 : 1,0,0,2,2$
$p_4 = 4 : 0,0,0,2,2$
$p_3 = 2 : 0,0,1,0,2$
  Groupsize: 4
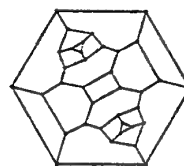
Nr.155 $v = 36$
$p_7 = 6 : 0,1,2,2,2$
$p_6 = 3 : 0,2,0,0,4$
$p_5 = 6 : 1,0,2,0,2$
$p_4 = 3 : 0,0,0,2,2$
$p_3 = 2 : 0,0,3,0,0$
  Groupsize: 12
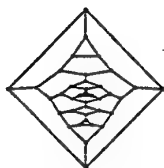
Nr.156 $v = 38$
$p_7 = 6 : 0,1,1,2,3$
$p_6 = 6 : 1,1,1,1,2$
$p_5 = 3 : 1,0,0,2,2$
$p_4 = 3 : 0,0,0,2,2$
$p_3 = 3 : 0,0,1,2,0$
  Groupsize: 6
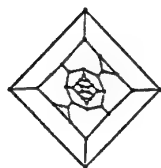
Nr.157 $v = 40$
$p_7 = 8 : 1,0,1,1,4$
$p_6 = 4 : 1,1,2,0,2$
$p_5 = 4 : 0,1,0,2,2$
$p_4 = 2 : 0,0,2,2,0$
$p_3 = 4 : 0,0,0,1,2$
  Groupsize: 4
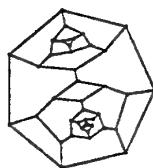
Nr.158 $v = 42$
$p_7 = 6 : 0,2,0,2,3$
$p_6 = 6 : 0,0,2,2,2$
$p_5 = 6 : 1,0,2,2,0$
$p_4 = 3 : 0,0,0,0,4$
$p_3 = 2 : 0,0,3,0,0$
  Groupsize: 12
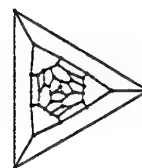
Nr.159 $v = 42$
$p_7 = 6 : 0,1,3,2,1$
$p_6 = 6 : 1,0,1,2,2$
$p_5 = 6 : 0,1,0,1,3$
$p_4 = 3 : 0,0,2,0,2$
$p_3 = 2 : 0,0,0,3,0$
  Groupsize: 6
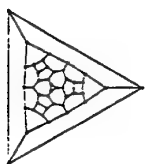
Nr.160 $v = 42$
$p_7 = 6 : 1,0,2,2,2$
$p_6 = 6 : 0,1,2,1,2$
$p_5 = 6 : 0,1,0,2,2$
$p_4 = 3 : 0,0,2,2,0$
$p_3 = 2 : 0,0,0,0,3$
  Groupsize: 12

26

4 List 3: Selected face-regular simple polyhedra with $b \geq 8$
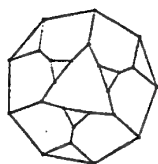
Nr.1 $v = 24$
$p_9 = 2 : 3,0,0,6,0,0,0$
$p_6 = 6 : 2,0,0,2,0,0,2$
$p_3 = 6 : 0,0,0,2,0,0,1$
Groupsize: 12
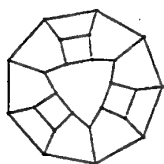
Nr.2 $v = 24$
$p_9 = 2 : 0,3,6,0,0,0,0$
$p_5 = 6 : 0,2,1,0,0,0,2$
$p_4 = 6 : 0,1,2,0,0,0,1$
Groupsize: 12
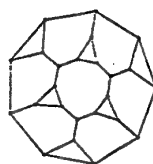
Nr.3 $v = 24$
$p_9 = 2 : 3,0,0,6,0,0,0$
$p_6 = 6 : 2,0,0,2,0,0,2$
$p_3 = 6 : 0,0,0,2,0,0,1$
Groupsize: 12
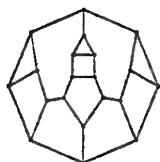
Nr.4 $v = 20$
$p_8 = 3 : 2,2,2,0,0,2$
$p_5 = 3 : 0,1,2,0,0,2$
$p_4 = 3 : 1,0,1,0,0,2$
$p_3 = 3 : 0,1,0,0,0,2$
Groupsize: 6
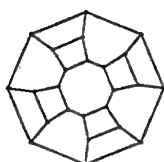
Nr.5 $v = 24$
$p_8 = 2 : 0,4,0,4,0,0$
$p_6 = 4 : 0,4,0,0,0,2$
$p_4 = 8 : 0,1,0,2,0,1$
Groupsize: 16
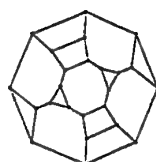
Nr.6 $v = 24$
$p_8 = 2 : 2,2,0,0,4,0$
$p_7 = 4 : 2,2,0,0,1,2$
$p_4 = 4 : 0,1,0,0,2,1$
$p_3 = 4 : 0,0,0,0,2,1$
Groupsize: 8
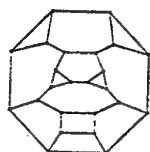
Nr.7 $v = 24$
$p_8 = 2 : 2,1,2,1,2,0$
$p_7 = 2 : 2,0,1,2,0,2$
$p_6 = 2 : 2,0,0,1,2,1$
$p_5 = 2 : 0,2,0,0,1,2$
$p_4 = 2 : 0,1,2,0,0,1$
$p_3 = 4 : 0,0,0,1,1,1$
Groupsize: 4
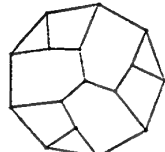
Nr.8 $v = 16$
$p_9 = 1 : 3,3,0,3,0,0,0$
$p_6 = 3 : 1,2,0,2,0,0,1$
$p_4 = 3 : 1,0,0,2,0,0,1$
$p_3 = 3 : 0,1,0,1,0,0,1$
Groupsize: 3
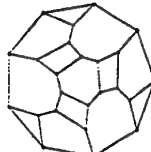
Nr.9 $v = 24$
$p_9 = 1 : 3,0,3,0,3,0,0$
$p_7 = 3 : 1,2,2,1,0,0,1$
$p_6 = 1 : 0,3,0,0,3,0,0$
$p_5 = 3 : 1,1,0,0,2,0,1$
$p_4 = 3 : 0,0,1,1,2,0,0$
$p_3 = 3 : 0,0,1,0,1,0,1$
Groupsize: 3

5 List 4: all 9 face-regular 4-valent polyhedra with $b = 4$

For the polyhedra in this list, the graph induced by the 4-gons is interesting: in Nrs 6,7,9 the graphs are two C_4, C_8 and the truncated Octahedron respectively.
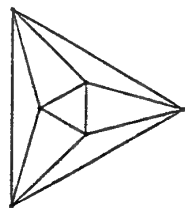
Nr.1 $v = 6$
$p_3 = 8$: 3
Groupsize: 48
Group: O_h
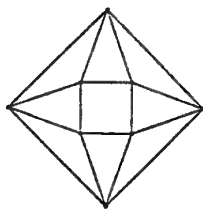
Nr.2 $v = 8$
$p_4 = 2$: 4,0
$p_3 = 8$: 2,1
Groupsize: 16
Group: D_{4d}
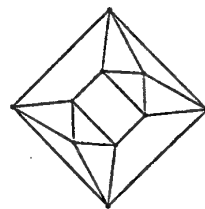
Nr.3 $v = 10$
$p_4 = 4$: 2,2
$p_3 = 8$: 2,1
Groupsize: 16
Group: D_{4h}
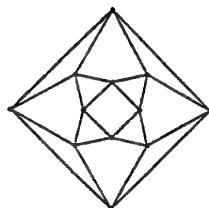
Nr.4 $v = 12$
$p_4 = 6$: 4,0
$p_3 = 8$: 0,3
Groupsize: 48
Group: O_h
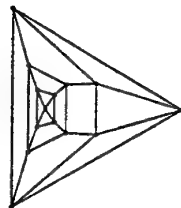
Nr.5 $v = 14$
$p_4 = 8$: 1,3
$p_3 = 8$: 2,1
Groupsize: 16
Group: D_{4h}
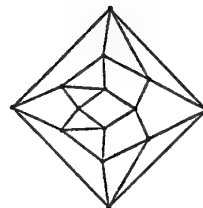
Nr.6 $v = 14$
$p_4 = 8$: 2,2
$p_3 = 8$: 1,2
Groupsize: 16
Group: D_{4h}
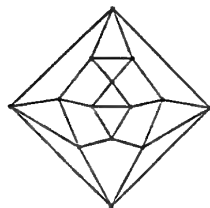
Nr.7 $v = 14$
$p_4 = 8$: 2,2
$p_3 = 8$: 1,2
Groupsize: 8
Group: D_{2d}
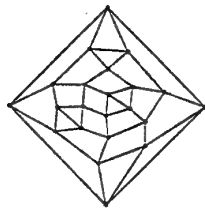
Nr.8 $v = 22$
$p_4 = 16$: 1,3
$p_3 = 8$: 1,2
Groupsize: 8
Group: D_{2d}
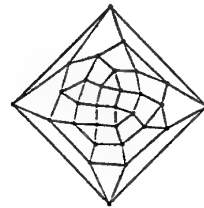
Nr.9 $v = 30$
$p_4 = 24$: 1,3
$p_3 = 8$: 0,3
Groupsize: 24
Group: O

6 List 5: all face-regular 4-valent polyhedra with $b = 5$ and up to 24 faces

Among the polyhedra of Lists 4 and 5 there are: the Octahedron, three semi-regular ones (4-, 5-gonal antiprisms and the Cuboctahedron) and three regular-faced (Nr. 3 of List 4 and Nrs.4 and 6 of List 5, which are the elongated square dipyramid, the pentagonal gyrobicupola and the pentagonal orthobicupola, having number 15, 31 and 30, respectively, in the list of 92 polyhedra in [Joh66]).

Nr.2 in list 5 is the Octahedron truncated and capped on 4 vertices of an induced C_4. Nr.3 is the elongated antiprism. Nr.5 is the dual rhombic Icosahedron (2-elongated 5-gonal antiprism).
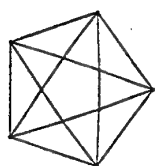
Nr.1 $v = 10$
$p_5 = 2$: 5,0,0
$p_3 = 10$: 2,0,1
Groupsize: 20
Group: D_{5d}
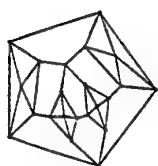
Nr.2 $v = 22$
$p_5 = 8$: 2,0,3
$p_3 = 16$: 2,0,1
Groupsize: 16
Group: D_{4h}
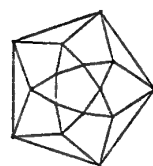
Nr.3 $v = 15$
$p_5 = 2$: 5,0,0
$p_4 = 5$: 4,0,0
$p_3 = 10$: 0,2,1
Groupsize: 20
Group: D_{5h}
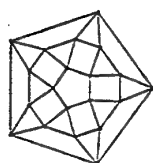
Nr.4 $v = 20$
$p_5 = 2$: 0,5,0
$p_4 = 10$: 3,0,1
$p_3 = 10$: 0,3,0
Groupsize: 20
Group: D_{5d}
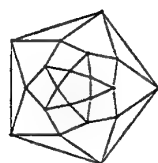
Nr.5 $v = 20$
$p_5 = 2$: 5,0,0
$p_4 = 10$: 2,2,0
$p_3 = 10$: 0,2,1
Groupsize: 20
Group: D_{5d}

Nr.6 $v = 20$
$p_5 = 2$: 0,5,0
$p_4 = 10$: 2,1,1
$p_3 = 10$: 1,2,0
Groupsize: 20
Group: D_{5h}

7 List 6: all 26 $6R_j$ Fullerenes

The polyhedra 1,3,6,12,16,18–21,24–26 of this list are face-regular. They are the polyhedra 23–34 of List 1, respectively.
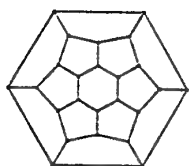
Nr.1 $v = 24$
$p_6 = 2 : 0,0,6,0$
Groupsize: 24
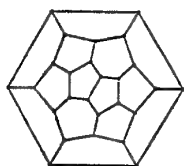
Nr.2 $v = 26$
$p_6 = 3 : 0,0,6,0$
Groupsize: 12
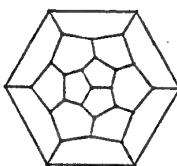
Nr.3 $v = 28$
$p_6 = 4 : 0,0,6,0$
Groupsize: 24
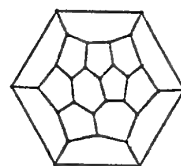
Nr.4 $v = 28$
$p_6 = 4 : 0,0,5,1$
Groupsize: 4
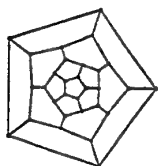
Nr.5 $v = 30$
$p_6 = 5 : 0,0,4,2$
Groupsize: 20
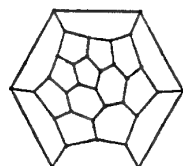
Nr.6 $v = 32$
$p_6 = 6 : 0,0,4,2$
Groupsize: 12
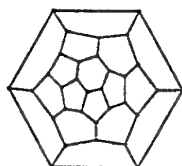
Nr.7 $v = 32$
$p_6 = 6 : 0,0,5,1$
Groupsize: 6
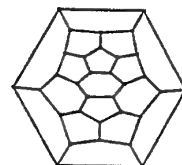
Nr.8 $v = 32$
$p_6 = 6 : 0,0,4,2$
Groupsize: 12
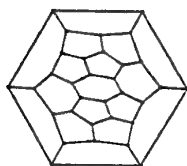
Nr.9 $v = 32$
$p_6 = 6 : 0,0,4,2$
Groupsize: 4
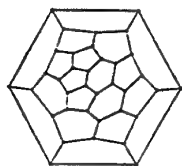
Nr.10 $v = 36$
$p_6 = 8 : 0,0,4,2$
Groupsize: 8
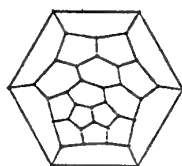
Nr.11 $v = 36$
$p_6 = 8 : 0,0,3,3$
Groupsize: 4
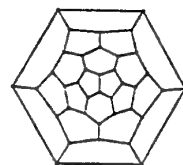
Nr.12 $v = 38$
$p_6 = 9 : 0,0,4,2$
Groupsize: 6
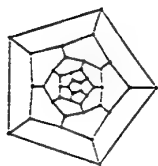
Nr.13 $v = 40$
$p_6 = 10 : 0,0,2,4$
Groupsize: 20
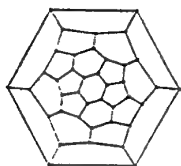
Nr.14 $v = 40$
$p_6 = 10 : 0,0,4,2$
Groupsize: 4
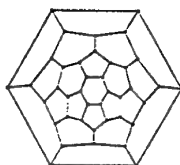
Nr.15 $v = 40$
$p_6 = 10 : 0,0,4,2$
Groupsize: 20
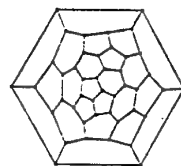
Nr.16 $v = 44$
$p_6 = 12 : 0,0,3,3$
Groupsize: 12

30
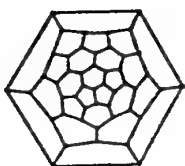
Nr.17 $v = 48$
$p_6 = 14 : 0,0,3,3$
Groupsize: 6
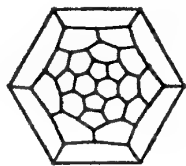
Nr.18 $v = 52$
$p_6 = 16 : 0,0,3,3$
Groupsize: 12
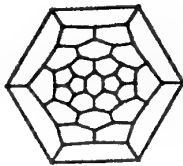
Nr.19 $v = 56$
$p_6 = 18 : 0,0,2,4$
Groupsize: 24
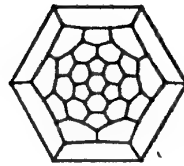
Nr.20 $v = 60$
$p_6 = 20 : 0,0.3,3$
Groupsize: 120
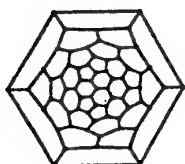
Nr.21 $v = 68$
$p_6 = 24 : 0,0,2,4$
Groupsize: 24
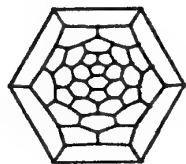
Nr.22 $v = 68$
$p_6 = 24 : 0,0,2,4$
Groupsize: 12
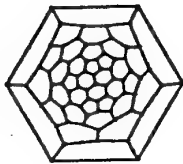
Nr.23 $v = 72$
$p_6 = 26 : 0,0,2,4$
Groupsize: 8
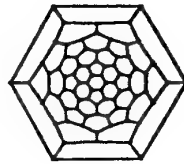
Nr.24 $v = 80$
$p_6 = 30 : 0,0,2,4$
Groupsize: 120
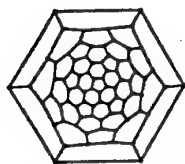
Nr.25 $v = 80$
$p_6 = 30 : 0,0,2,4$
Groupsize: 20
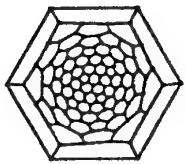
Nr.26 $v = 140$
$p_6 = 60 : 0,0,1,5$
Groupsize: 60

31

Face-regular k-valent bifaced polyhedra

In this section we will study the case where faces of exactly two sizes $a < b$ occur.

Bifaced polyhedra and similar concepts are well studied, e.g. in [Mal70], [GM66], [GZ74]. [Go75]. [Go77].[Za80]. [JT84]. [JT90], [JJ89]. [Ow84]. [Ow86]. [JO95].

Clearly. in this case the f-graph from the proof of Theorem 1 is the K_2, so we get the equation $p_b = p_a \frac{f(a,b)}{f(b,a)}$.

By using $kv = 2e = ap_a + bp_b$ and the Euler formula $v - e + (p_a + p_b) = 2$, we get

(*) $p_a = \frac{4k}{(2k+2a-ak)-(bk-2b-2k)\frac{f(a,b)}{f(b,a)}}$.

We will use the following notation for operations on polyhedra:

tetrakis: the tetrakis of a polyhedron is obtained by putting a pyramid on 4-gonal faces.

4-triakon: the 4-triakon of a polyhedron is obtained by partitioning each triangle into a ring of 3 4-gons by putting a vertex in the middle and connecting it to the midpoint of every edge in the boundary. ·

5-triakon: a 5-triakon of a polyhedron is obtained by partitioning each hexagon into a ring of 3 pentagons by putting a vertex in the middle and connecting it to the midpoint of every second edge in the boundary. (An example of two different face-regular bifaced polyhedra, coming both as a 5-triakon of the truncated Octahedron, is given in Remark 1 after the Theorem 5 below.)

Theorem 2 *For $k > 3$ there is only one infinite series of face-regular (a, b)-polyhedra, that is the Antiprisms $APrism_b$ for any $b > 3$.*

Apart from this, all face-regular k-valent (a, b)-polyhedra have $(k; a) = (4; 3)$ and $b = 4, 5. 6$;

They are:

$b = 4$: *7 polyhedra, given as Nrs 3-9 in List 4;*

$b = 5$: *the Icosidodecahedron and Nr 2 in List 5 (the tetrakis of the Octahedron, truncated on all but two opposite vertices);*

$b = 6$: *the tetrakis of the (fully) truncated Octahedron.*

Theorem 2 will follow from the following 4 Lemmata:

Remark

Theorem 2 shows that the largest 4-valent face-regular (a, b)-polyhedra have 30 vertices (i.e. 32 faces) and $a = 3$. They have $(p_3, b) \in \{(8, 4), (20, 5), (24, 6)\}$ and are Nr 9 in List 4, the Icosidodecahedron, the tetrakis of the truncated Octahedron, respectively. The largest 3-valent face-regular $(3, b)$-polyhedron also has 32 faces. It is the fully truncated Dodecahedron with $(p_3, b) = (20, 10)$.

As we will see below, all three largest 3-valent face-regular polyhedra have 140 vertices. They are unique largest $(4, b)$-polyhedron (the 4-triakon of the truncated Dodecahedron; so $p = (p_4 = 60, p_{15} = 12)$) and both largest $(5, b)$-polyhedra: a 5-triakon of the truncated Icosahedron;so $p = (p_5 = 60, p_{10} = 12)$) and the fullerene $C_{140}(I)$ (the truncation of the dual snub Dodecahedron on all 12 5-valent vertices; so $p = (p_5 = 12, p_6 = 60)$).

Lemma 1 *The only possibilities for $(k; a, b)$ are $(5; 3, 4), (4; 3, b), (3; 3, b), (3; 4, b)$ and $(3; 5, b)$.*

In all other cases the denominator in (*) will be non-positive even for $\frac{f(a,b)}{f(b,a)} = \frac{1}{b}$, the smallest possible value.

Lemma 2 *The case $(k; a, b) = (5; 3, 4)$ is not possible, so there is no face-regular 5-valent polyhedron.*

Proof.

The denominator in (*) is positive only for $(f(a, b), f(b, a)) = (1, 4), (1, 3)$.

In the first case any 4-gon is surrounded by 12 3-gons, which implies neighbouring 4-gons in the next layer – a contradiction. In the case $(1, 3)$, any pair of adjacent 4-gons is surrounded by 16 3-gons, so the next layer contains a 4-gon with 2 4-gonal neighbours – again a contradiction.

□

In the following lemma we will exclude some of the theoretically possible parameters for $k = 4$.

Lemma 3 *All cases for $(k; a) = (4; 3)$ not being contained in the following list are impossible:*

a): $b = 4$: *the 8 polyhedra Nrs 2-9 in List 4;*

b): $b = 5$: *the Icosidodecahedron with $f(3, 5) = 3, f(5, 3) = 5, v = 30$;*

c): $b > 3$: *the infinite class of antiprisms $APrism_b$;*

d): $b > 3$: $(f(3, b), f(b, 3) = (1, b - 3), (p_3, p_b) = (8b - 24, 8), v = 8b - 18$;

e): $(f(3, b), f(b, 3) = (1, b - 2), (p_3, p_b) = (4b - 8, 4), v = 4b - 6$.

Proof. For $b > 4$ the denominator in (*) is positive only if

1) $(f(3, b), f(b, 3)) \in \{(1, b - 3), (1, b - 2), (1, b - 1), (1, b)\}$, or

2) $b \in \{5, 6, 7\}$ and $(f(3, b), f(b, 3)) = (2, b)$, or

3) $b \in \{5, 6\}$ and $(f(3, b), f(b, 3)) = (2, b - 1)$, or

4) $b = 5$ and $f(3, b), f(b, 3)) \in \{(2, 3), (3, 4), (3, 5)\}$.

The subcase $(1, b - 1)$ in case 1) is not possible, because otherwise $p_3 = \frac{8(b-1)}{3}$, $p_b = \frac{8}{3}$. The subcases $(1, b), (1, b - 3), (1, b - 2)$ of the case 1) are, respectively, the cases c, d and e of Lemma 3.

Cases 2 and 3 are not possible, because we get 3 3-gons on 3 consecutive edges of each b-gon; so, the 3-gonal neighbour of the 3-gon in the middle, will be adjacent to 2 3-gons, a contradiction.

The subcase $(3, 4)$ of 4 is not possible, since the 3-gonal neighbours of 2 adjacent 5-gons containing a vertex of the intersection, would share an edge.

In the subcase $(2, 3)$ of 4, all 3 triangles neighbouring a pentagon in a row would imply a triangle neighbouring 2 other ones, so assume we have a 5-gon and

33

3 neighbouring 3-gons not all in a row. But then one of the 5-gonal neighbours has all 3-gonal neighbours in a row – again a contradiction.

The remaining subcase (3.5) of 4 is case b of Lemma 3.

□

Lemma 4 *The cases d and e of Lemma 3 are realized only by polyhedron Nr.3 in List 4 and the tetrakis of suitably truncated Octahedra, given in Theorem 2.*

Proof.

In both cases all 3-gons are organized in 4-cycles, surrounded by b-gons, since otherwise we will get $APrism_b$. In case e of Lemma 3 the number of edges between two b-gons is $\frac{p_b(b-f(b,3))}{2} = 4$. So, the only possibility is $b = 4$ and Nr 3 of List 4 is unique realization (it is the dual of the Octahedron, truncated on 2 opposite vertices). In case d of Lemma 3, the number of $(b-b)$-edges is 12. This implies $b \in \{4, 5, 6\}$ and we get the tetrakis of 3 suitably truncated Octahedra, the first one being Nr 5 of List 4 (the elongated Nr 3 of the List).

□

Theorem 3 *All face-regular cubic $(3, b)$-polyhedra have $b \leq 10$.*

They are 14 special truncations of the Tetrahedron, the Cube and the Dodecahedron:

- *the 1- and 4-truncated Tetrahedron;*

- *$2(b-4)$-truncated Cubes (one for $b \in \{5, 7, 8\}$ and two for $b = 6$);*

- *$4(b-5)$-truncated Dodecahedra (one for $b = 6, 9, 10$ and two for $b = 7, 8$).*

Proof.

Due to the 3-connectedness of polyhedra we get $f(3, b) = 3$ for all cubic $(3, b)$-polyhedra. So each triangle is isolated and $3p_3 \leq v = p_3(2 + \frac{6}{f(b,3)}) - 4$. Together with equality (*) and $p_3 > 0$ we get $f(b, 3) \geq b - 5$ and $f(b, 3) \leq min(5, \frac{b}{2})$. So $b \leq 10$.

Actually this is a result by Malkevitch ([Mal70]) for general (that is: not only face-regular) cubic $(3, b)$-polyhedra.

The remaining possibilities for $b > 6$ are $(b, f(b, 3); v) \in \{(10, 5; 60), (9, 4; 52), (8, 3; 44), (7, 2; 36), (8, 4; 24)$ and $(7, 3; 20)$. The first 4 cases are realized by truncations of the Dodecahedron giving only one polyhedron in the first two cases and two non-isomorphic polyhedra in the others. The last 2 cases are realized by truncations of the Cube (giving a unique polyhedron in every case). For $b \leq 6$ all wanted polyhedra are Nrs 4,5 and 10-13 of List 1.

Nrs 1-3 of List 2 are the 6-truncated Cube and two 8-truncated Dodecahedra. For $b > 6$ there remain 3 (3,8)-polyhedra, one (3,9)- and one (3,10)-polyhedron.

□

Remark

If we do not require 3-connectedness in Theorem 3, more graphs exist, e.g. one for every $b \geq 9$, divisible by 3 with $p_b = 2, p_3 = \frac{2b}{3}, v = \frac{4b}{3}$.

Theorem 4 *There is only one infinite family of cubic $(4, b)$-polyhedra, that is Prism₄ for any $b \geq 3$;*

The finite families are:

1) *Two 80-vertex $(4, 7)$-polyhedra. coming as the truncation of the dual Rombicuboc-tahedron or dual Miller's solid (its twist) on all 18 4-valent vertices;*

2) *14 polyhedra, coming from those of Theorem 3 by the 4-triakon decoration; they have $(b, v) \in \{(15, 140), (13.116), (11, 92), (9, 68), (7, 44), (12, 56), (10, 44), (8, 32), (6, 20); (9, 28) (6, 14)\}$. There are exactly 2 non-isomorphic polyhedra for the 3rd, 4th and 8th case and one for the others.*

3) *3 polyhedra (besides the prism) for $b = 5$ and 6 for $b = 6$ not covered by previous cases: Nrs 7-9 and 17-22 of List 1;*

4) *2 44-vertex $(4, 7)$-polyhedra Nrs 5,6 of List 2 (coming by suitable doubling of all 6 isolated 4-gons in Nrs 21,20 of List 1);*

5) *the 80-vertex $(4,8)$-polyhedron (coming by suitable doubling of all 12 isolated 4-gons of the truncation on all 12 4-valent vertices of the dual of the unique 18-vertex 4-valent $(3,4)$polyhedron given below, the 3-gons of which are organized into 2 isolated ones and 3 isolated pairs);*
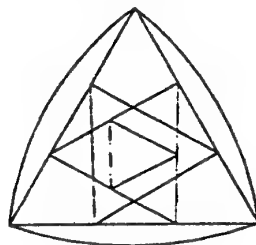
Proof.

Possible values for $f(4, b)$ are $2, 3$ and 4. The case $f(4, b) = 2$ is possible only if the 4-gons form a ring (giving a prism) or isolated 3-rings of 4-gons, which is exactly case 2 of Theorem 4.

If $f(4, b) = 4$, all 4-gons are isolated. So $4p_4 \leq v$. The equality (*) and $p_4 > 0$ imply $2(b - 6) < f(4, b) \leq min(3. \frac{b}{2})$. So $b \leq 7$ and the only possibility for $b = 7$ is $f(b, 4) = 3, p_4 = 18, p_7 = 24; v = 80$. It is exactly case 1 of Theorem 4.

In the remaining case $f(4, b) = 3$, the 4-gons are organized into isolated adjacent pairs. So, $6\frac{24}{2} \leq v$ and, using (*) and $p_4 > 0$, we get $\frac{3(b-6)}{2} < f(b, 4) \leq min(5, \frac{3b}{5})$, which implies $b \leq 9$. Moreover. the only possible values for $(b, f(b, 4))$ are $(9, 5)$, $(8, 4)$, $(7, 2)$, $(7, 3)$ and $(7, 4)$. The last subcase is not possible, because it gives $p_4 = \frac{48}{5}$. The subcase $(7, 3)$ gives $v = 44$. A computer search gave that it is exactly case 4 of the theorem. The remaining subcases leave 3 possibilities: $(b, f(b, 4); p_4, p_b; v) \in \{(7, 2; 24, 36; 116), (8, 4; 24, 18; 80), (9, 5; 60, 36; 188)\}$. The first of them can easily be removed by a geometric argument. For the last one there are 8 vertices contained only in 9-gons. It is easy to show that only one out of two possible ways in which the pairs of 4-gons can neighbour a 9-gon containing such a vertex can occur. Using this, geometric arguments give a contradiction when trying

to construct the polyhedron. The middle one is realized by the polyhedron of case 5 of the theorem. The uniqueness follows from its construction.

Clearly, all remaining wanted polyhedra have $b \leq 6$ and so they are covered by List 1: it gives the last entry of case 3 in Theorem 4.

$\square$

Remark

The largest face-regular cubic $(4,6)$- and $(5,6)$-polyhedra are also (just like in case 1 of Theorem 4) the dual tetrakis or dual snub Cube and dual pentakis (putting pyramids on all 5-gonal faces) of the snub Dodecahedron.

Theorem 5 *There is only one infinite family of face-regular cubic $(5,b)$-polyhedra with $b \geq 6$:*

$Barrel_b$ (two b-gons, separated by two rows of b 5-gons) for any b.
The finite families are:

1) *12 $(5,6)$-polyhedra: Nrs 23-34 in List 1;*

2) *a unique 92-vertex $(5,7)$-polyhedron, organized into concentric 3-, 15- and 12-ring of 5-gons, separated by 6-, 9- and 3-ring of 7-gons (as given below);*

3) *a unique 44-vertex $(5,7)$-polyhedron (Nr 9 in List 2) and 3 polyhedra with $f(5,b) = 2$, also given below: unique 140-vertex $(5,10)$-polyhedron, unique 56-vertex $(5,8)$-polyhedron and unique 92-vertex $(5,8)$-polyhedron with $(p_5, p_8; f(8,5)) = (36, 12; 6)$.*
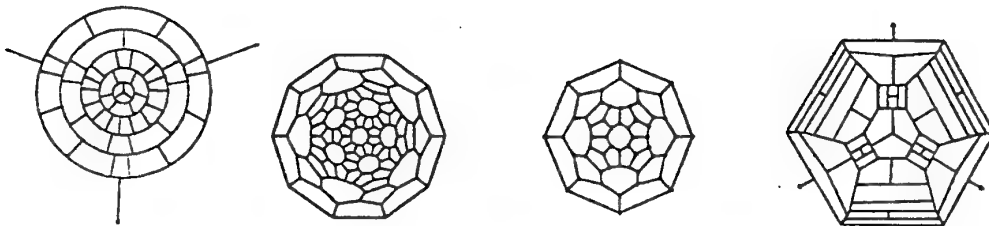
Proof.

The case $f(5,b) = 1$ gives exactly $Barrel_b$. If $f(5,b) = 5$, then all 5-gons are isolated; so $5p_5 \leq v$. So (*) and $p_5 > 0$ imply $5(b-6) < f(b,5) \leq 3$ giving $b < 7$.

If $f(5,b) = 4$, then all 5-gons are organized into isolated pairs; so $8\frac{p_5}{2} \leq v$. Again we get $4(b-6) < f(b,5) \leq 3$ and $b < 7$.

The case $f(5,b) = 3$ has 5-gons organized in disjoint rings. Let t denote the number of 3-rings among them, so $3p_5 + t \leq v$. The same count as above, gives $3(b-6) < f(b,5) \leq min(5, \frac{3b(t+4)-18}{t+16})$. So $b = 7$ is the only possibility for $b > 6$. In the case $b = 7$ we have either $f(7,5) = 4$ and $t \leq 20$, or $f(7,5) = 5$ and $t \leq 2$. The first subcase gives $p_5 = 48$, $p_7 = 36$ and it should be 164-vertex $(5,7)$-polyhedron with all 48 5-gons organized into isolated rings. V.P.Grishukhin (private communication) established, case by case, non-existence in this subcase. The second subcase is $f(7,5) = 5$, i.e. 7-gons also should be organized in isolated rings. We get $p_5 = 30$, $p_7 = 18$; $v = 92$. So 5- and 7-gons should be organized in

36

concentric alternating rings and only two vertices belong to three faces of the same type. It is easy to show that case 2 of the theorem is the unique possibility for such a polyhedron.

All possibilities with $b = 6$ are covered by List 1 (it is case 1 of the theorem). The only remaining case is $f(5, b) = 2, b > 6$. Using (*), we get $2(b - 6) < f(b, 5) \leq b$. Clearly, $p_b = p_5 \frac{f(5,b)}{f(b,5)} = \frac{24}{f(b,5) - 2(b-6)}$ and $v = 2p_5 + 2p_b - 4 = \frac{8b + 20f(b,5)}{f(b,5) - 2(b-6)}$. So all possibilities with positive integer v are given by $f(b, 5) = b - i$, $b \leq 11 - i$ for $0 \leq i \leq 4$. In subcase $f(b, 5) = b$ the b-gons are all isolated. It is easy to see that b should be even and that for $b \in \{8, 10\}$ the only possibilities are the 140- and 56-vertex polyhedra in case 3 of the theorem. In subcase $f(b, 5) = b - 1$ an attempt to construct the structures easily gives the impossibility for $b \in \{9, 8\}$ and unicity (the 44-vertex polyhedron of case 3 of the theorem) for $b = 7$. The impossibility of cases $(b, f(b, 5); v) = (7, 5; 52), (7, 4; 68)$ was checked with the help of a computer. The remaining 4 cases should be $(5, b)$-polyhedra with $f(5, b) = 2$, having $(b, f(b, 5); p_5, p_b; v) = (7, 3; 36, 24; 116), (8, 5; 60, 24; 162), (8, 6; 36, 12; 92), (9, 7; 84, 24; 212)$. The third possibility is realized, uniquely, by the 92-vertex polyhedron of the case 3 in the theorem. It and the non-existence in other 3 subcases can be checked by following easy way. Let us fix a b-gon $A_0 = (1, 2, ..., b)$ and, without loss of generality, suppose that other b-gon, say A_1, adjacent to A_0 by edge (1,2). In all 4 subcases it is not possible, that A_0 was adjacent to a 5-gon by the edge (2,3) and to a b-gon by the edge (3,4), because then this 5-gon will have 3, not 2 b-gons as neighbours. Now, consider the situation when A_0 is adjacent to 5-gons by (2,3) and by (3,4), but not by (4,5). A try to construct a polyhedron, respecting our conditions on $f(5, b), f(b, 5)$, will continue uniquely and lead to impossibility always except the subcase 3. (Most difficult is the situation when all $b - f(b, 5)$ b-neighbours of the original b-gon A_0 are adjacent to it in a row: by (1,2), (2,3) and so on.)

□

Remark 1

The polyhedron Nr 24 of List 1 (i.e. the fullerene $F_{28}(T_d)$) and the 3-rd and 2-nd polyhedron in case 3 of Theorem 5 come by a *5-triakon* decoration of the (fully) truncated Tetrahedron, Octahedron and Icosahedron, respectively.

The 4-th polyhedron of the case 3 of Theorem 5 (92-vertex (5,8)-polyhedron) comes from face-regular fullerene $F_{56}(T_d)$ (Nr 29 of List 1) by following decoration of (all 6) its hexagons, having two adjacent 5-gons, being adjacent on opposite edges: put some "H" with sides parallel to above opposite edges, so that hexagon will be partitionned into 4 pentagons. Above face-regular fullerene comes itself by a 5-triakon decoration (another one, than one, producing above 56-vertex (5,8)-polyhedron) of a face-regular 24-vertex (4,6)-polyhedron: truncated Octahedron.

Remark 2

Above Therems 3, 4 and 5 give together the following classification:

Besides of two infinite families, $Prism_b$ and $Barrel_b$, there are exactly 57 3-valent face-regular bifaced polyhedra. With respect of the number v of vertices, they have the face-sizes (a, b) as follows.

for $v = 140$: (4,15), (5,6), (5,10);
for $v = 116$: (4,13);
for $v = 92$: two (4,11), (5,7), (5,8);
for $v = 80$: two (4,7), (4,8), two (5,6);

37

for $v = 68$: two (4,9), (5,6);
for $v = 60$: (3,10), (5,6);
for $v = 56$: (4,6), (4,12), (5,6), (5,8);
for $v = 52$: (3,9);
for $v = 44$: two (3,8), three (4,7), (4,10), (5,6), (5,7);
for $v = 38$: (5,6);
for $v = 36$: two (3,7);
for $v = 32$: (3,8), two (4,6), two (4,8), (5,6);
for $v = 28$: (3,6), (4,9), (5,6);
for $v = 24$: (4,6), (5,6);
for $v = 20$: (3,7), (4,6);
for $v = 16$: two (3,6), (4,5);
for $v = 14$: (4,5), three (4,6);
for $v = 12$: (3,5), (3,6), (4,5).

References

[Ba90] L. Balke, *Diskontinuierliche Gruppen als Automorphismengruppen von Pflasterungen*, Diplomarbeit Universität Bonn (1990), Bonner Mahematische Schriften Nr. 211.

[BDr96] A. Dress and G. Brinkmann, *Phantasmagorical Fulleroids, MATCH* (1996) 87–100.

[De90] O. Delgado, *Die automatische Konstruktion periodischer Pflasterungen*, Diplomarbeit Universität Bielefeld (1990).

[DeDez98] O. Delgado-Friedrichs and M. Deza, *More icosahedral fulleroids*, Report LIENS 98-6, Ecole Normale Supérieure Paris (1998), DIMACS Series in Discrete Mathematics and Theoretical Computer Science.

[DGr97a] M.Deza and V.P.Grishukhin, *A zoo of l_1-embeddable polyhedra*, Bull. Inst. Math. Acad. Sinica **25** (1997) 181–231.

[DGr97b] M.Deza and V.P.Grishukhin, *A zoo of l_1-embeddable polyhedra II*, Research Report LIENS 97-9, Ecole Normale Supérieure Paris , Proc. of Int. Conference on Algebra and Combinatorics (1997).

[DGr97c] M.Deza and V.P.Grishukhin, *Face-regular bifaced polyhedra*, Research Report LIENS 97-10, Ecole Normale Supérieure Paris (1997), Journal of Statistical Planning and Interference (1998), Special Issue in honour of S.S.Shrikhande, ed. by V.Tonchev.

[Go75] P.R.Goodey, *Hamiltonian circuits in polytopes with even sided faces*, Israel Journal of Mathematics **22** (1975) 52–56.

[Go77] P.R.Goodey, *A class of Hamiltonian polytopes*, Journal of Graph Theory **1** (1977) 181–185.

[Gr67] B. Grünbaum, *Convex Polytopes*, John Wiley & Sons, Ltd. (1967)

[GM66] B.Grünbaum and T.Motzkin, *The number of hexagons and the simplicity of geodesics on certain polyhedra*, Canadian Journal of Mathematics **15** (1963) 744–751.

[GZ74] B.Grünbaum and J.Zaks. *The existence of certain planar maps*, Discrete Mathematics **10** (1974) 93–115.

[JJ89] S.Jendrol and E.Jucovic, *On quadrangular convex 3-polytopes with at most two types of edges*, Discrete Mathematics **78** (1989) 297–305.

39

[JT84] S.Jendrol and M.Tkac, *On the simplicial 3-polytopes with only two types of edges,* Discrete Mathematics **48** (1984) 229–241.

[JT90] S.Jendrol and M.Tkac. *Convex 3-polytopes with exactly two types of edges,* Discrete Mathematics **84** (1990) 143–160.

[JO95] S.Jendrol and P.J.Owens. *Pentagonal 3-polytopal graphs with edges of only two types and shortness parameters,* Discrete Mathematics **137** (1995) 251–263.

[Joh66] N.W.Johnson, *Convex polyhedra with regular faces,* Canadian J. Math. **18** (1966) 169–200.

[Ma71] P. Mani, *Automorphismen von polyedrischen Graphen,* Math. Ann. **192** (1971) 279–303.

[Mal70] J.Malkevitch, *A survey of 3-valent polytopes with 2 types of faces* in: Combinatorial Structures and their Applications, Proceedings of Calgary Conference 1969, ed. R.Guy and al. (1970) 255–256.

[Ow84] P.J.Owens, *Non-Hamiltonian simple 3-polytopes with only one type of face besides triangle,* Annals of Discrete Mathematics 20 (1984) 241–251.

[Ow86] P.J.Owens, *Simple 3-polytopal graphs with edges of only two types and shortness coefficients,* Discrete Mathematics **59** (1986) 107–114.

[Za80] J.Zaks, *Non-Hamiltonian simple 3-polytopes having just two types of faces,* Discrete Mathematics **29** (1980) 87–101.

Some relations for the ground state energy and helium diatomic molecules

S. Kilić and L. Vranješ
Faculty of Natural Science, University of Split,
21000 Split, Croatia

April 21, 1999

Abstract

It is shown quite generally that ground state energy of two atoms in infinite space, interacting *via* spherical potential which depends only on the distance between particles, is the lowest in two dimensions. Using variational procedure, binding energies of helium diatomic molecules, in infinite and restricted space, are obtained as well. The results derived for helium atoms are in accordance with the lemma.

1   Introduction

Many physical phenomena in nature are related to the behaviour of the small number of particles. Among them, in low temperature physics are superconductivity, superfluidity and Bose-Einstein condensation. Special interesting and important cases are systems in which particles are helium atoms: helium liquids, helium films, liquid drops, atoms in cavities in solid matrices and in nanotubes.

The consideration of small systems begins with study of two atoms. They can be located in both restricted and unrestricted space: in 3 dimensions (3 D), 2 dimensions (2 D) and 1 dimension (1 D). Of course, real physical world has been occuring in finite 3 dimensional space. In making models of different physical situations we are led to consider 2 D and 1 D space. In such circumstances many physical effects are dominant in corresponding dimension.

In this paper, in Sec. I, we prove a general lemma. It relates ground state energies of two particles in 1, 2 and 3 dimensions in infinite space. It is assumed that particles interact via spherical potential depending only on the distance between them. In Sec. II, using variational procedure and employing the newest potential of the interaction between helium atoms [1], the ground state energies of helium molecules are obtained. The consistency with the lemma is demonstrated.

2 Relations between ground state energies in different dimensions

We consider two particles which interact *via* a spherically symmetrical potential $\hat{V}(\vec{r}_1, \vec{r}_2)$, in one, two and three dimensional space. The Hamiltonian of the system in the relative coordinates reads

$$\hat{H} = -\frac{\hbar^2}{2\mu}\Delta + \hat{V}(|\vec{r}_1 - \vec{r}_2|), \tag{1}$$

where $\mu = \frac{m_1 m_2}{m_1 + m_2}$ is reduced mass of the particles, m_1 and m_2 are the masses of the particles. In the ground state only the "radial" part of the Hamiltonian is important and the operator Δ has the form

$$\Delta_1 = \frac{\partial^2}{\partial r^2}, \qquad\qquad in\ 1D \tag{2}$$

$$\Delta_2 = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}, \qquad in\ 2D \tag{3}$$

$$\Delta_3 = \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}, \qquad in\ 3D. \tag{4}$$

Inequalities between energies in different dimensions may be obtained by variational *ansatz*

$$E_n \leq \frac{\int \Psi_n^\star \hat{H}_n \Psi_n\ r^{n-1} dr\ d\Omega^n}{\int \Psi_n^\star \Psi_n\ r^{n-1} dr\ d\Omega^n}, \tag{5}$$

where $n = 1, 2, 3$ denotes the dimension of physical space and $d\Omega^1 = 1$, $d\Omega^2 = 2\pi$ and $d\Omega^3 = 4\pi$.

As we study the ground state and having in mind the symmetry of the system, it is useful to write the trial wave functions in the form

$$\Psi_1(r) = \Psi_{10}(r)$$
$$\Psi_2(r) = \frac{1}{\sqrt{r}}\Psi_{20}(r)$$
$$\Psi_3(r) = \frac{1}{r}\Psi_{30}(r). \tag{6}$$

Introducing trial wave functions in the variational *ansatz* (5) one finds

$$E_1 \leq \frac{1}{I_1}\left[-\frac{\hbar^2}{2\mu}\int_0^\infty dr\Psi_1\frac{d^2}{dr^2}\Psi_1 + \int_0^\infty dr\Psi_1^2\,V(r)\right] \tag{7}$$

$$E_2 \leq \frac{1}{I_2}\left[-\frac{\hbar^2}{2\mu}\int_0^\infty dr\Psi_{20}\{\frac{d^2}{dr^2}\Psi_{20} + \frac{1}{4r^2}\Psi_{20}\} + \int_0^\infty dr\Psi_{20}^2\,V(r)\right] \tag{8}$$

$$E_3 \leq \frac{1}{I_3}\left[-\frac{\hbar^2}{2\mu}\int_0^\infty dr\Psi_{30}\frac{d^2}{dr^2}\Psi_{30} + \int_0^\infty dr\Psi_{30}^2\,V(r)\right], \tag{9}$$

3

where the normalization integrals read $I_n = \int_0^\infty dr \Psi_{n0}^2$, n=1,2,3.

The relations (6), (7) and (8) are general. Assuming that Ψ_1 is the eigenfunction in one dimensional case and taking $\Psi_{20} = \Psi_1$, from (6) and (7) it follows

$$E_2 < E_1 - \frac{1}{I_1} \frac{\hbar^2}{2\mu} \int_0^\infty dr \frac{1}{4r^2} \Psi_{20}^2; \qquad (10)$$

it means $E_2 < E_1$.

If it is supposed $\Psi_{30} = \Psi_{20}$, where Ψ_2 is the eigenfunction in 2 D, then from (7) and (8) one finds

$$E_3 < E_2 + \frac{1}{I_2} \frac{\hbar^2}{2\mu} \int_0^\infty dr \frac{1}{4r^2} \Psi_{20}^2. \qquad (11)$$

On the other hand, if Ψ_3 is the eigenfunction in 3 D and $\Psi_{20} = \Psi_{30}$, then

$$E_3 > E_2 + \frac{1}{I_3} \frac{\hbar^2}{2\mu} \int_0^\infty dr \frac{1}{4r^2} \Psi_{30}^2. \qquad (12)$$

The last two inequalities may be joined

$$E_2 + \frac{1}{I_3} \frac{\hbar^2}{2\mu} \int_0^\infty dr \frac{1}{4r^2} \Psi_{30}^2 < E_3 < E_2 + \frac{1}{I_2} \frac{\hbar^2}{2\mu} \int_0^\infty dr \frac{1}{4r^2} \Psi_{20}^2. \qquad (13)$$

From the above relation it follows $E_2 < E_3$. In a similar consideration from (6) and (8) follows $E_3 = E_1$. In this way it is proved that binding energy of two interacting particles is the lowest in 2 D. The result is independent on the statistic of the particles.

3   Ground state energy of diatomic helium molecules

In order to describe physical systems that contain helium, many potentials between atoms have been obtained. One of the best is *ab initio* SAPT potential by Korona et al. [1]; its enlarged forms by Janzen and Aziz [2] are SAPT1 and SAPT2 which comprise retardation effects. Since the SAPT potential is so precise, it is expected that the effect of retardation forces could be examined experimentally. It reads

$$V(r) = \epsilon V^*(r) \qquad (14)$$

$$V^*(r) = A e^{-\alpha r + \beta r^2} - B \sum_{n=3}^{8} f_{2n}(b, r) \frac{c_{2n}}{r^{2n}}, \qquad (15)$$

4

where

$$f_{2n}(b,r) = 1 - e^{-br} \sum_{k=0}^{2n} \frac{(br)^k}{k!}$$ (16)

and

$$\epsilon = 10^8 mK, \qquad c_6 = 0.03207856 \,\text{Å}^6$$
$$A = 20.7436426 \qquad c_8 = 0.08680214 \,\text{Å}^8$$
$$B = 3.157765 \qquad c_{10} = 0.31625734 \,\text{Å}^{10}$$
$$\alpha = 3.56498393 \,\text{Å}^{-1} \qquad c_{12} = 1.57407624 \,\text{Å}^{12}$$
$$\beta = -0.22141687 \,\text{Å}^{-2} \qquad c_{14} = 10.31938196 \,\text{Å}^{14}$$
$$b = 3.68239497 \,\text{Å}^{-1} \qquad c_{16} = 86.00126516 \,\text{Å}^{16}.$$

As first let us calculate binding energy of two helium atoms in infinite space. For the ground state we found that a good analytic form of the functions (6) in all dimensions is

$$\Psi_{0i}(r) = \exp\left[-\left(\frac{a}{r}\right)^\gamma - sr\right],$$ (17)

where i=1,2,3; a, γ and s are variational parameters and of course have different minimization values in 3 D (1 D) and 2 D. The same form of pair correlations in 3 D has recently been used by Bruch [3] to examine the properties of boson trimers. In 2 D, we use the form employed in the paper [4] and which provides a slight improvement over a variational wave function introduced in Ref. [5]. Binding energy and parameters are obtained in minimization procedure. The results are shown in Table I. In order to estimate our variational calculation, and compare the results, corresponding numerical solutions of Schroedinger eq. are presented for HFD-B3-FCI1 [6] and SAPT potentials as well.

Now, as second, we concentrate on two helium atoms confined by a hard-walled spherical potential in 3 D and circular in 2 D. As it was demonsrated in the paper [4], good variational wave functions of the ground state are

$$\Psi_{03}(r;d) = \Psi_{03}(r)j_0(\pi r/d)$$ (18)

in 3 D, and

$$\Psi_{02}(r;d) = \Psi_{02}(r)J_0(2.404826r/d)$$ (19)

in 2 D. d is the diameter of the sphere and of the circle. j_0 is the spherical Bessel function and J_0 is the zeroth-order Bessel function. As in infinite space, the ground state energy of the non-interacting system must be subtracted. The energy of two free particles is $\frac{C_i}{d^2}$, i=2,3, where $C_2 = \hbar^2 (2.404826)^2/2\mu$ in 2 D and $C_3 = \hbar^2\pi^2/2\mu$ in 3 D. The results for d=50 Å and d=100 Å are presented in Table II.

5

4 Discussion

Let us mention that only helium 4 dimer in 3 D has been observed experimentally [7-9] up to now.

As it is seen in Table 1, binding energies for all helium molecules are consistent with the lemma. Moreover both lighter molecules are not bound at all in 3 D. Two particles may be kept in 2 D space by an external potential. It can be realized, for example, in a space between two close, paralel big plates. Similarly interior of a long and thin cylinder may represent 1 D space. Of course these "confining" external potentials are not included in binding energies cited in Table 1.

Since in restricted geometry (in our case sphere and cylinder) external potentials are included partly, the lemma can not be valid. Of course it is correct in this case as well, if parameters of the geometry (for instance in our case diameter of sphere or cylinder) are much biger than the effective range of the interaction potential. Such behaviour can be recognized in Table 2.

From the "exact" numerical solution of Schroedinger eq. [4] we know that all combinations of two helium atoms are bound in finite space (in above sense); the same is in infinite space except two atoms of ^{3}He and one atom ^{3}He and one atom ^{4}He which are not bound in 3 D. Let us notice that our trial function in the case of $(^3$He$)_2$ is not good enough to reproduce binding in 2 D in both infinite and finite space. As comparision with numerical solution shows, it is quite good for other cases.

It seems that an interior of a cylinder is a form which could be the easyest to realize in an experiment. Although we haven't solved this problem theoretically, the main energetic characteristics are given by our spherical-models in 3 D and 2 D.

Finally let us mention that our calculation in finite space is approximative one. Namely we assumed that the center of mass of two particles was located in the center of space symmetry. It was shown in Ref. [4] that this approximation gives general features of considered systems.

References

[1] T. Korona *et al.*, J. Chem. Phys. **106**, 5109 (1997).

[2] A. R. Janzen and R. A. Aziz, J. Chem. Phys. **107**, (1997).

[3] L. Bruch, J. Chem. Phys. **110**, 2410 (1999).

[4] S. Kilić, E. Krotscheck and R. Zillich, accepted in J. Low Temp. Phys. (1999).

[5] S. Kilić and S. Sunarić, Fizika **11**, 225 (1979).

[6] R. A. Aziz, A. R. Janzen, and M. R. Moldover, Phys. Rev. Lett. **74**, 1586 (1995).

[7] F. Luo, G.C. McBane, G. Kim, C.F. Giese and W.R. Gentry, J. Chem. Phys. **98**, 3564 (1993).

[8] F. Luo, C.F. Giese and W.R. Gentry, J. Chem. Phys. **104**, 1151 (1996).

[9] W. Schollkopf and JP. Toennies, J. Chem. Phys. **104**, 1155 (1996).

5 ACKNOWLEDGEMENTS

Table 1: Binding energies in infinite space (in mK) of helium molecules in 2 D and for dimer $(^4He)_2$ in 3 D (second line), derived by numerical solving Schroedinger eq. and in variational procedure; variational values are in round brackets; parameters: a (in Å), γ (dimensionless) and s (in Å^{-1}), are shown for the SAPT potential only. Note that our variational wave function is not flexible enough to predict a bound state of the $(^3He)_2$ dimer in 2 D and that molecules $(^3He)_2$ and $^3He -^4 He$ are not bound in 3 D.

Molecule	HFD-B3-FCI1[a]	SAPT[b]	a	γ	s
$(^4He)_2$	-39.4 (-37.7)	-40.7 (-39.93)	2.758	4.408	0.047
	-1.559 (-1.480)	-1.871 (-1.762)	2.737	4.49	0.012
$(^3He)_2$	-0.016	-0.02			
$^3He -^4 He$	-4.0 (-3.21)	-4.3 (-3.51)	2.761	4.173	0.011

[a] Ref. [6]
[b] Ref. [1]

8

Table 2: Binding energies (in mK) of helium molecules in a sphere (3 D, first line) and in a circle (2 D, second line) derived in variational procedure for the SAPT potential; the diameter of both confinements are d=50 Å and d=100 Å; parameters: a (in Å), γ (dimensionless) and s (in Å^{-1}), are shown for d=50 Å.

Molecule	50	100	a	γ	s
$(^4He)_2$	-138.713	-40.650	2.753	4.41	-0.013
	-61.660	-52.133	2.767	4.36	0.02
$(^3He)_2$	-67.086	-10.191	2.782	3.91	-0.058
	73.159	14.827	2.798	3.87	-0.029
$^3He - ^4He$	-94.354	-19.936	2.774	4.10	-0.042
	16.718	-7.264	2.794	4.04	-0.011

9

Subtle Relationships exist among apparently unconnected properties

Lionello Pogliani

Dipartimento di Chimica, Universita' della Calabria, 87030 Rende (CS), Italy

Abstract

Similar molecular connectivity terms are capable to model many different properties of quite different classes of compounds, like : alkanes, amino acids, purines, pyrimidines and inorganic salts. Modeled properties are : the pH at the isoelectric point, pI, the specific rotations, the solubility, the side-chain molecular volume, the crystal densities of amino acids, the solubility of purines and pyrimidines, the solubility of amino acids plus purines and pyrimidines, the lattice enthalpies of metal halides, the unfrozen water content of amino acids plus metal chlorides, and the motor octane numbers of alkanes. The internal formal similarity of these different higher-level molecular connectivity descriptors, which are derived by a trial-and-error procedure from a medium-sized set of 8 molecular connectivity indices or a subset of it is striking. Nearly all of them are dominant terms, that is, they are descriptors which are unable to further enhance the description when they are used in combinations with other descriptors.

2

Subtle Relationships exist among apparently unconnected properties

Lionello Pogliani

Dipartimento di Chimica, Universita' della Calabria, 87030 Rende (CS), Italy

Abstract

Similar molecular connectivity terms are capable to model many different properties of quite different classes of compounds, like : alkanes, amino acids, purines, pyrimidines and inorganic salts. Modeled properties are : the pH at the isoelectric point, pI, the specific rotations, the solubility, the side-chain molecular volume, the crystal densities of amino acids, the solubility of purines and pyrimidines, the solubility of amino acids plus purines and pyrimidines, the lattice enthalpies of metal halides, the unfrozen water content of amino acids plus metal chlorides, and the motor octane numbers of alkanes. The internal formal similarity of these different higher-level molecular connectivity descriptors, which are derived by a trial-and-error procedure from a medium-sized set of 8 molecular connectivity indices or a subset of it is striking. Nearly all of them are dominant terms, that is, they are descriptors which are unable to further enhance the description when they are used in combinations with other descriptors.

INTRODUCTION

Recently [1-5] the modeling of different physicochemical properties of molecules or materials has allowed to discover quite interesting and powerful descriptors, the molecular connectivity terms, $X = f(\chi)$, which are based on graph theoretical invariants, the molecular connectivity indices. These are second- or higher-level descriptors, derived by a trial-and-error procedure on medium set of molecular connectivity indices. It is, thus, possible to detect that interesting relationships exist among many and different molecular and material properties. Achieved descriptions show that physicochemical properties have a common root in the chemical graph or pseudograph representation of a molecule, which are the starting 'material' to construct the molecular connectivity indices, the most widely used invariants in QSPR/QSAR studies [6] (see references therein). Interesting, the form of these terms, used to describe quite different properties of different classes of molecules is rather homogeneous.

In this short paper we will review the different properties which can be modeled by the aid of molecular connectivity terms, check the advantages, over the normal molecular connectivity indices, of a modeling with such terms and indirectly stress how higher-order invariants derived from chemical graphs and pseudographs can be the common descriptors of quite different properties of different classes of compounds.

Modeled classes of compounds and properties are : for amino acids (AA) : pH at the isoelectric point, pI, side-chain molecular volume, V, specific rotations, SR_L (SR_D for D-AA have just opposite values), crystal densities, CD, solubility, S ; for purines and pyrimidines : solubility, S ; for a mixed class of amino acids plus purines and pyrimidines: solubility, S[AA+PP] ; for metal halides : lattice enthalpies, $\Delta H_L^{\varnothing}$; for a mixed class of amino acids plus metal chlorides : unfrozen water content, UWC, and, finally, for alkanes : the motor, MON, octane number.

METHOD

Molecular connectivity terms are derived with a trial-and error composition procedure performed on a set of 8 molecular connectivity indices : $\{D, D^v, {}^0\chi, {}^0\chi^v, {}^1\chi, {}^1\chi^v, \chi_t, \chi_t^v\}$. Derivation of these indices from the corresponding graphs or pseudographs (which allow multiple connections and loops) is a straightforward, and has already been explained elsewhere [6, 1, 2]. Sometimes this medium-sized set can be restricted to a subset of optimal χ descriptors, derived with a combinatorial technique, that is, a technique that searches the entire combinatorial space spanned by the 8 indices, which means 255 combinations. The choice of 8 main χ indices alone, is done to ease the combinatorial problem both at the level of the choice of

the best index or indices, and at the level of the trial-and-error procedure to derive the the corresponding terms. Interesting, the molecular connectivity terms can, sometimes, further be combined with χ indices to derive a mixed linear combination with improved modeling capability, even if, normally molecular connectivity terms are dominant descriptor and do not allow to derive any linear combination with improved modeling power. Formally, the molecular connectivity terms have the following form

$$ X\left(a,b,\chi_i,\chi_j,\chi_k,\chi_l\right) = \frac{\chi_i + a\chi_j}{\chi_k + b\chi_l} \qquad (1) $$

Where a and b are optimization parameter, that can be positive, negative or zero. Indices χ_i, χ_j, χ_k and χ_l are indices of the given molecular connectivity set. Further, each index can be elevated to a positive, negative or zero power. Normally, the trial-and-error search technique is quite straightforward and convergence is easily reached [1-5]: just start with an optimal index, add to it the next one, after optimization of this index, back-optimize the previous one, then construct the fraction and act in the same way, and, finally, a and b coefficients as well as powers are added and optimized.

Indices of set $\{\chi\}$ are based on the degree δ_i and valence degree, δ_i^v (for the valence molecular connectivity indices), of each vertex i of a molecular graph and pseudograph respectively, where for degree is meant the number of connections incident to that vertex. Pseudographs allow for multiple connections and loops, that is, self-connections, that contribute twice to δ_i^v. For components of metal halides following definition, $\delta_i^v = Z_i^v / (Z_i - Z_i^v - 1)$ has been chosen [6], where Z_i^v is the number of valence electrons and Z is the atomic number of the corresponding atom.

Aim of the modeling is to describe a set of properties by the aid of a linear or multi-linear relationship, $P = \Sigma_i m_i X_i$, where $X = \chi$ represents a special case, m ranges from 0 to n and for n = 0, we have, $X_0 = X^0 \equiv 1$. Values of m_i constants are obtained with a linear least-squares procedure. Normally, with X terms we have i = 0,1, that is, a simple linear relationship. Negative meaningless results for modeled properties can be avoided using $|\Sigma_i m_i X_i|$, where bars stand for absolute values, and use of such an algorithm normally improve the description. If, instead, negative values are experimentally justified, as it is the case for SR_D and SR_L of amino acids, then bars are omitted. Used statistics to check the validity of molecular connectivity indices or terms are : the quality parameter, $Q = r / s$, the F ratio, $F = fr^2/[(1-r^2)v]$, where, r = correlation

4

coefficient, s = standard deviation of the estimate, f = degrees of freedom, v = number of χ or/and X parameters, the utility, $u_k = |c_k/s_k|$ of every descriptor, inclusive the unitary $X^0 \equiv 1$ term of the constant parameter, as well as the average utility $<u> = \Sigma u_k/m$ of the m descriptors.

DISCUSSION

The experimental data as well the molecular connectivity indices of the different classes of compounds are collected in refs 1-5 as well as in refs. 7, and 8.

Modeling the pH at the isoelectric point, pI, of amino acids

An appropriate and unusual term for this property of 21 amino acids which can be guessed even before any trial-and-error procedure is started, is the term given by eq. 2 [8]. In fact, considerations concerning the importance of the number of functional basic and/or acidic groups in amino acids are critical for this kind of property. In this term, $\Delta n = n_A - n_B$, where n_A = n° of acidic groups (2 for Asp and Glu, and 1 for all others), n_B = n° of basic groups (2 for Lys and His, 3 for Arg, and 1 for all others), and $n_T = 3$, that is, the total number of functional groups ; notice that for $n_T = 2$, $\Delta n = 0$.

$$^1X_{pI} = \frac{\chi}{^0\chi^v}\left(1 + \frac{\Delta n}{n_T}\right) \qquad (2)$$

For $\chi = {}^0\chi^v$ we obtain the following interesting modeling (where the rationale for notation is : for $\chi = D^v \rightarrow X_{pI} \equiv {}^DX^v$, for $\chi = {}^0\chi^v$, $X_{pI} \rightarrow {}^0X^v$, and so on).

$\{{}^0X^v\}$: $Q = 2.12$, $F = 267$, $r = 0.966$, $s = 0.46$, $<u> = 22.4$

Now, found term is rather trivial as it is nothing else than $(1 + \Delta n /n_T)$, even if the following combination of 4 terms with a better Q, r and s values is less trivial

$\{{}^0X^v, {}^0X, {}^0X^v, {}^1X\}_{pI}$: $Q = 2.53$, $F = 95$, $r = 0.980$, $s = 0.39$, $<u> = 7.9$

But here, the good F value of the single descriptor together with its excellent, $u = (16.3, 28.4)$, utility vector are lost. In fact, the single utilities, with the exception of $u(X^0)$, this last combination are rather deceiving, with $u = (3.1, 2.8, 4.7, 2.8, 26.3)$. Now, a deeper trial-and-error search reveals the existence of the following not at all trivial term

5

$$^2X_{pl} = \frac{(^1\chi^v)^{0.5} - 180\chi_t^v}{D}\left(0.04 \cdot \chi_t^v + \frac{\Delta n}{n_T}\right) \qquad (3)$$

with an interesting improvement in modeling power : $Q = 3.412$, $F = 693$, $r = 0.987$, $s = 0.29$, $\langle u \rangle = 58$, $u = (26, 90)$, $C = (77.99429, 5.75382)$. Not only the improvement in F and u is more than expected, but, further, this term is a highly dominant term (or 'dead-end' term), as it doesn't allow any better combination with any other index or term.

Modeling the side-chain molecular volume, V, of L-amino acids

The description of this property for 18 amino acids offers some interesting clues about use, form and ratings of connectivity terms. The best χ and LCCI (linear combination of χ indices) are: $\{^0\chi^v\}$: $Q = 0.25$, $F = 691$, $r = 0.989$, $s = 4.0$, $\langle u \rangle = 15$, and $\{D, ^0\chi^v, ^1\chi^v\}$: $Q = 0.43$, $F = 688$, $r = 0.997$, $s = 2.3$, $\langle u \rangle = 4.8$. Now, the following term $^1X_V = (D^v)^{3.5}/^0\chi^v$, found by a trial-and-error procedure [8], is a quite poor descriptor with, $Q = 0.031$, $F = 11$, and $r = 0.632$. But, together with $^0\chi^v$ shows the following surprising modeling and utility: $\{^0\chi^v, ^{3.5}X\}$: $Q = 0.424$, $F = 989$, $r = 0.996$, $s = 2.4$, $\langle u \rangle = 16$, $u = (34, 5.3, 7.4)$. We might really wonder if somewhere there is not a better term apt to describe this property in a more satisfactory way. A deeper search discovers the following term, which show: $Q = 0.438$, $F = 2109$, $r = 0.996$, $s = 2.3$, $\langle u \rangle = 32$, and $u = (46, 17)$

$$^2X_V = \frac{(D^v)^{1.3} + (^0\chi)^{2.1}}{D^v - 0.7 \cdot D} \qquad (4)$$

and whose correlation vector is : $C = (18.1182, -52.5871)$. The statistical improvement of this last term is quite impressive, and, clearly, it is a dominant 'dead-end' term.

Modeling the specific SR rotations of L-amino acids

An optimal term for the specific rotation SR_L of n=16 L-AA in aqueous solution, can be found with a trial-and-error composition procedure applied to the optimal restricted set, $\{D, ^0\chi, \chi_t\}$, which rates [3,4] : $Q = 0.088$, $F = 41.2$; $r = 0.955$, $s = 10.8$, $\langle u \rangle = 7.2$. Found term, X_{SR}, shown in eq. 5, when $a = 7$, rates : $Q = 0.04$, $F = 30$; $r = 0.830$, $s = 19.0$, $\langle u \rangle = 5.6$. This term

can be also be used, with $C_L = -C_D$, to describe the specific rotations SR_D of the corresponding D-AA, which differ from SR_L in their sign only.

$$X_{SR}(a) = \frac{^0\chi}{(D + a\chi_i)} \qquad (5)$$

This quality of this term becomes evident when it is compared with the quality of the single best $\{\chi_i\}$ index, which has, $Q = 0.014$, $F = 3.2$, $r = 0.43$, $s = 30.4$, $<u> = 2.1$. Now, a combinatorial search on the following set, $\{D, D^v, {^0\chi}, {^0\chi^v}, {^1\chi}, {^1\chi^v}, \chi_t, \chi_t^v, X_{SR}(1)\}$, where X_{SR} has here, $a = 1$, allows to find the following optimal combination

$$\{{^1\chi}, X_{SR}(1)\} : Q = 0.100, F = 79.5, r = 0.961, s = 9.6, <u> = 10.6$$

A search with $X_{SR}(7)$ replacing $X_{SR}(1)$ achieves a worse description, with $\{{^1\chi}, \chi_t, X_{SR}(7)\}$: $Q = 0.097$, $F = 50$, $r = .962$, $s = 9.9$, $<u> = 8.1$. Thus, chosen modeling vectors for SR_L are: $\mathbf{X} = ({^1\chi}, X_{SR}, X^0)$, $\mathbf{C_L} = (26.28495, 965.8255, -545.67)$, $\mathbf{u} = (7.6, 12.3, 11.8)$, where in the last vector we can notice the good utility of every term.

The fact that the found family of molecular connectivity terms is not able to derive a good enough single descriptors oblige us to deepen the trial-an-error search around a term composed of 4 descriptors. The search ends up with the following dominant satisfactory term, which being a dead-end term does not allow the construction of any improved multilinear description

$$^2X_{SR} = \frac{^0\chi - (\chi_i^v)^{0.3}}{D^{0.8} + 0.2(\chi_i)^{0.02}} \qquad (6)$$

that shows the following statistics and $\mathbf{C}$ vector : $Q = 0.084$, $F = 112$, $r = 0.943$, $s = 11.2$, $<u> = 10.7$, $\mathbf{u} = (10.6, 10.9)$, $\mathbf{C} = (573.114, -430.56)$.

The modeling of Crystal density, CD, of Amino Acids

The crystal densities, CD, of 10 amino acids cannot satisfactorily be modeled with any χ index, in fact, the single index description of this property is quite poor[4]

7

$\{^0\chi^v\}$: $Q = 3.43$, $F = 3.9$, $r = 0.57$, $s = 0.2$, $<u> = 5.4$

Now, a trial-and-error search for a better descriptor finds the following molecular connectivity term, whose improved description, nevertheless, is rather deceiving

$$\left\{ \, ^1X_{CD} = \frac{^0\chi^v}{(^1\chi + 1)} \, \right\} : \quad Q = 5.86, \, F = 11.2, \, r = 0.76, \, s = 0.13, \, <u> = 5.3 \qquad (7)$$

The following more convoluted term of eq. 8 enhances the description consistently but always in an unsatisfactory way, with $Q = 7.91$, $F = 20.4$, $r = 0.848$, $s = 0.1$, $<u> = 5.4$, $u = (4.5, 6.4)$, the correlation vector being, $C = (-0.50967, 4.81717)$. Both terms are dominant 'dead-end' terms.

$$^2X_{CD} = \frac{(^0\chi^v)^{1.2} + 1.8 \cdot {}^0\chi}{(^1\chi)^{0.8} - 1.3(\chi_t)^{2.1}} \qquad (8)$$

Modeling the solubility of amino acids, purines and pyrimidines and of [AA + PP]

We will now model the solubility (g per Kg of H_2O) of $n = 20$ amino acids, the solubility of 23 purines and pyrimidines, and finally the solubility of a mixed class of 43 amino acids plus purines and pyrimidines. The modeling of the solubility of the two separate classes of compounds, AA or PP, has already quite satisfactorily been achieved by the aid of supramolecular reciprocal [9] and supramolecular squared [10] molecular connectivity terms, respectively. Supramolecular connectivity indices are obtained multiplying (χ_t and χ_t^v indices are normally divided, as they are total reciprocal indices) the molecular connectivity indices by an association parameter, a, to take into account supposed or detected association phenomena in solution [11]. To model the entire class of [AA+PP] compounds the introduction of the following set of supraindices is mandatory [8]

$$\{aD \cdot \chi_t^v, \, aD^v \cdot \chi_t^v, \, a^0\chi \cdot \chi_t^v, \, a^0\chi^v \cdot \chi_t^v, \, a^1\chi \cdot \chi_t^v, \, a^1\chi^v \cdot \chi_t^v, \, \chi_t \cdot a^{-1}, \, \chi_t^v \cdot a^{-1}\} \qquad (9)$$

where $a = 8$ for Pro, 2 for Ser, Hyp, and Arg, and 1 for the remnant amino acids, while for PP we have $a = 4$ for 7PTp, 2 for 7Etb, ETp, and Cf, 1.5 for 7Itp and 1 for the remnant PP. To notice that the formulation of the given set is slightly different from the one already given in a

preceding paper [8]. The supraindices of this set , to avoid to burden the resulting equations, will be renamed into

$$\{{}^{D}S, {}^{D}S^{v}, {}^{0}S, {}^{0}S^{v}, {}^{1}S, {}^{1}S^{v}, S_{t}, S_{t}^{v}\} \tag{10}$$

We will start looking for a term to model the 20 amino acids (no Cys, but with Hyp) using the given supraindices as our final aim is to model the entire heterogeneous class of [AA+PP]. Dominant 'dead-end' term of eq. 11 can achieve a quite satisfactory modeling, with, $Q = 0.020$, $F = 1007$, $r = 0.991$, $s = 49.3$, $<u> = 21$, $u = (32, 11)$

$$X_{S}(AA) = \frac{{}^{D}S^{v} - {}^{D}S^{0.3}}{(\chi_{t} + 350 \cdot \chi_{t}^{v})^{0.7}} \tag{11}$$

whose correlation vector is $C = (4650.56, -162.218)$.

The modeling of the 23 purines and pyrimidines, can also be achieved by the somewhat different dominant 'dead-end' term of eq. 12, with $Q = 0.282$, $F = 4005$, $r = 0.997$, $s = 3.5$, $<u> = 33$, $u = (64, 3.1)$, and $C = (11.6437, -1.53752)$

$$X_{S}(PP) = \frac{{}^{1}S^{v} + ({}^{0}S)^{1.1}}{(S_{t} - 0.002)^{1.2}} \tag{12}$$

The only deceiving note in this modeling is the poor utility of the unitary term of the correlation vector. Notice that we are here modeling S in g per Kg of H_2O, while in our preceding modeling of S(PP) solubility has been modeled g in 100 ml H_2O [10].

Well, let us now model the entire [AA + PP] class of compounds, and let us start this modeling with terms, $X_S(AA)$ and $X_S(PP)$. Their modeling power of the whole heterogeneous class of compounds is

$X_S(AA)$: $Q = 0.020$, $F = 1079$, $r = 0.982$, $s = 50.0$, $<u> = 24$, $u = (33, 15)$

$X_S(PP)$: $Q = 0.010$, $F = 297$, $r = 0.937$, $s = 91.0$, $<u> = 9.9$, $u = (17, 2.6)$

The term for the solubility of amino acids is quite good, and the term for the solubility of PP is not at all deceiving. In fact the first term for amino acids is very similar, in form and modeling power, to the following term, which has been found with a trial-and-error procedure over the entire class [AA + PP]

$$X_S(AA+PP) = \frac{(^D S^v)^{1.1} - (^D S)^{1.1}}{(\chi_t + 10^3 \cdot \chi_t^v)^{0.7}} \qquad (13)$$

This term rates : $Q = 0.021$, $F = 1199$, $r = 0.983$, $s = 47$, $<u> = 23$, $u = (35, 11)$, and the correlation vector is : $C = (7958.87, -100.596)$. This term is also a dominant term even if in combination with index D it shows a somewhat improved Q statistics

$\{X_S(AA+PP), D\} : Q = 0.024$, $F = 779$, $r = 0.987$, $s = 42$, $<u> = 17$, $u = (38, 3.6, 7.8)$

Modeling the Lattice Enthalpy of Metal Halides

The optimal modeling of the lattice enthalpies of 20 metal halides by molecular connectivity indices and by the reciprocal molecular connectivity indices, $^1R^v = (^1\chi^v)^{-1}$, and $^{1.5}R^v = (^1\chi^v)^{-1.5}$, seems rather satisfactory, especially at the level of two-index linear combination [7,8]

$\{^0\chi^v\}$: $Q = 0.015$, $F = 45$, $r = 0.846$, $s = 57.3$, $<u> = 17.2$

$\{D^v, ^0\chi^v\}$: $Q = 0.033$, $F = 115$, $r = 0.965$, $s = 29.0$, $<u> = 19.4$

$\{^1R^v\}$: $Q = 0.019$, $F = 72$, $r = 0.895$, $s = 48.1$, $<u> = 29.8$

$\{^1R^v, ^{1.5}R^v\}$: $Q = 0.038$, $F = 147$, $r = 0.972$, $s = 25.9$, $<u> = 24.2$

Now, a trial-and-error procedure finds a term, shown in eq. 14, based on two types of molecular connectivity indices, that shows the following statistics : $Q = 0.037$, $F = 281$, $r = 0.969$, $s = 24.4$, $<u> = 41$, $u = (17, 65)$, with correlation vector, $C = (1911.76, 623.102)$

$$X_H = \frac{(D^v)^{0.5} + 0.2}{D^v + 4.2 \cdot ^0\chi^v} \qquad (14)$$

10

This dominant term allows for a somewhat improved description with $^1\chi^v$

$\{X_H, {}^1\chi^v\}$: Q =0.041, F = 173, r = 0.976, s = 24.0, <u> = 15, u = (10.2, 2.2, 32)

but the improvement in Q, does not suffice to overrule the worsening of F and u, thus, X_{II} can also be considered a 'dead-end' term

Modeling the unfrozen water content UWC of amino acids and inorganic salts

The modeling of this property for 5 metal chlorides and 8 amino acids by normal χ indices is quite deceiving [7] while the following trial-and-error term of eq. 15 [3,4] is quite satisfactory, with Q = 2.79, F = 328, r = 0.984, s = 0.35, <u> = 12.8, and u = (18, 7.6)

$$X_{UWC} = \frac{{}^1\chi^v}{(D^v - {}^0\chi^v)} \qquad (15)$$

A somewhat better modeling can be achieved if the absolute values of X_{UWC} are considered, that is, $| X_{UWC} |$. In this case we obtain the following interesting statistics, Q = 3.14, F = 417, r = 0.987, s = 0.3, <u> = 12.8, u = (20, 5.0), with C = (1.83423, 0.55209). While X_{UWC} allows for an enhanced Q/u-combination with χ_t^v with following values, Q = 3.11, F = 203, <u> = 9 ; $| X_{UWC} |$ term is a strict 'dead-end' term allowing no enhanced Q and u combinations.

Modeling the motor MON octane number of alkanes

A trial-and-error search for a molecular connectivity term with the $\{D, {}^0\chi, {}^1\chi, \chi_t\}$ set of alkanes, that do not have any valence molecular connectivity indices, discovers the term of eq.16, with the following statistics : Q = 0.085, F = 146, r = 0.916, s = 10.8, <u> = 19.5, u = (12.1, 27)

$$X_{MON} = \frac{({}^0\chi \cdot \chi_t)^{0.1} + (D)^{1.3}}{({}^0\chi - 1.5 \cdot {}^1\chi)^{1.2}} \qquad (16)$$

The correlation vector being, $C = (-1.6714, 121.02)$ This term offers the possibility, through a combinatorial search, to find a linear combination, with a consistently enhanced Q, r and s statistics, but a worse F and u statistics

$\{X_{MON}, D, {}^0\chi, {}^1\chi\}$: $Q = 0.129$, $F = 85$, $r = 0.965$, $s = 7.9$, $<u> = 7.5$, $u = (4.8, 5.6, 5.4, 5.4, 4.1)$

CONCLUSION

Molecular connectivity terms whose general type has been described in the method section seem to be the subtle thread that ties together the many different properties of different classes of compounds. These terms, based on molecular connectivity indices are the last step in the search of a formally common descriptor for different properties of different molecular structures. Not only the similarity among the many different terms is striking, but the fact that normally a single term is sufficient for a good description, make us confident that they represent a powerful tool to derive and infer physicochemical information about molecules. The trial-and-error procedure to derive molecular connectivity terms is easier than the combinatorial technique used to detect the optimal set of best molecular connectivity indices for a multilinear description of a property. This procedure is in some way similar to a forward selection combinatorial technique, where only the next best index is chosen, to which a a back-optimization step has been added. Further, this procedure allows to find more than one term that accomplishes its task in an optimal way. For example, for UWC it is possible to find also the following term, $|[(0.1({}^1\chi)^{0.5}+0.6{}^1\chi^v)/(D^v - {}^0\chi^v)]|$, which rates $Q = 3.12$, $F = 410$, $r = 0.987$, $s = 0.3$, $<u> = 12.8$, and $u = (20, 5.3)$, and the modeling of the solubility of [AA + PP], that can be accomplished with the same term used to describe the solubility of amino acids underlines the good 'adaptability' of these terms. And, as the search for terms belongs to the more general search for new invariants, it is then not at all odd here to cite W. Ostwald, who was very aware about the importance of invariants in science : '*The significance of a law of nature .... is the finding of an invariant, that is to say, a quantity which remains unchanged even when all the other determining elements vary within the possible limits imposed by the law. Thus, we perceive that the historical development of scientific concept is ever associated with the discovering and working out of such invariants, in them we behold the milestones which mark the track traversed by human knowledge.*'

REFERENCES

(1) Pogliani, L. Modeling biochemicals with leading molecular connectivity terms. *Med.Chem.Res.* **1997**, *7*, 180-393.

(2) Pogliani, L. Modeling properties of biochemical compounds with connectivity terms. *Amino Acids*, **1997**, *13*, 237-255.

(3) Pogliani, L. Properties of molecular connectivity terms and physicochemical properties. *J.Mol.Struct.(Theochem)* **1999**, *466*, xx-xx (to be published).

(4) Pogliani, L. Modeling properties with higher-level molecular connectivity descriptors. *J.Chem.Inf.Comput.Sci.* **1999**, *39*, 104-111.

(5) Pogliani, L. Modeling with semiempirical molecular connectivity terms. *J.Phys.Chem.* **1999**, *103*, 1598-1610.

(6) Kier, L.B.; Hall, L.H. *Molecular connectivity in structure-activity analysis.* Wiley : New York, 1986.

(7) L. Pogliani, Modeling enthalpy and hydration properties of inorganic compounds. *Croat.Chim.Acta* **1997**, *70*, 803-817.

(8) L. Pogliani, Modeling with special descriptors derived from a medium-sized set of connectivity indices. *J.Phys.Chem.* **1996**, *100*, 18065-18077.

(9) Pogliani, L. A strategy for molecular modeling of a physicochemical property using a linear combination of connectivity indices. *Croat.Chim.Acta* **1996**, *69*, 95-109.

(10) Pogliani, L. Modeling purines and pyrimidines with the linear combination of connectivity indices – molecular connectivity "LCCI-MC" method. *J.Chem.Inf.Comput.Sci.* **1999**, *36*, 1082-1091.

(11) Pogliani, L. Molecular modeling by linear combinations of connectivity indices. *J.Phys.Chem.* **1996**, *99*, 925-937.

Fullerenes as tilings of surfaces

M. Deza[a], P.W. Fowler[b], A. Rassat[c] and K.M. Rogers[b]

[a] Laboratoire d'Informatique de l'Ecole Normale Supérieure, 45 rue d'Ulm, 75230 Paris, France

[b] School of Chemistry, University of Exeter, Stocker Road, Exeter EX4 4QD, UK

[c] Department of Chemistry, Ecole Normale Supérieure, rue Lhomond, 75005 Paris. France

Abstract

If a fullerene is defined as a finite trivalent graph made up solely of pentagons and hexagons, embedding in only four surfaces is possible: the sphere, torus, Klein bottle and projective (elliptic) plane. The usual spherical fullerenes have 12 pentagons, elliptic fullerenes 6, and toroidal and Klein-bottle fullerenes none. Klein-bottle and elliptic fullerenes are the antipodal quotients of centrosymmetric toroidal and spherical fullerenes, respectively. Extensions to infinite systems (plane fullerenes, cylindrical fullerenes and space fullerenes) are indicated. Eigenvalue spectra of all four classes of finite fullerenes are reviewed. Leapfrog fullerenes have equal numbers of positive and negative eigenvalues, with 0, 0, 2 or 4 eigenvalues zero for spherical, elliptic, Klein-bottle and toroidal cases, respectively.

Introduction

The discovery of the fullerene molecules and related forms of carbon such as nanotubes has generated an explosion of activity in chemistry, physics and materials science, which is amply documented elsewhere [1-4]. In chemistry, the 'classical' definition is that a fullerene is an all-carbon molecule in which the atoms are arranged on a pseudo-spherical framework made up entirely of pentagons and hexagons, which therefore necessarily includes exactly 12 pentagonal rings. 'Non-classical' extensions to include rings of other sizes have been considered (e.g. ref. [5]) and may be competitive in energy with the classical fullerenes in some ranges of nuclearity (e.g. ref. [6]). The present paper is concerned with a generalisation in a different direction: what fullerenes are possible if a fullerene is a finite trivalent map with only 5- and 6-gonal faces embedded in any surface (i.e. a 2-manifold in the mathematical sense)? This seemingly much larger concept leads to a small number of well defined extensions to the class of spherical fullerenes, actually three in number. Of

1

these only the toroidal fullerenes are likely to have direct experimental relevance (indeed observation of a toroidal 'fullerene crop circle' has already been reported [7]) but all three extensions are useful in placing physical fullerenes in a wider mathematical context and are considered in this light here. A more mathematical treatment of the concept of the extended fullerenes and their further generalisation to higher dimensional spaces is given elsewhere [8].

Classification of finite fullerenes

Define a fullerene in the wider sense as a finite, trivalent map on a surface and consisting of only 5-gonal and 6-gonal faces. Each such object has n vertices, e edges and f faces of which f_5 are pentagons and f_6 hexagons. Infinite analogues of fullerenes will be considered in a later section.

The Euler characteristic χ is defined as the number

$$\chi = n - e + f \tag{1}$$

which for a trivalent graph (hence having $2e = 3n$) made up of pentagons and hexagons (and hence $2e = 5f_5 + 6f_6$) is

$$\chi = f_5/6. \tag{2}$$

For a surface in which a fullerene in this extended sense can be embedded, the number χ is therefore a non-negative integer. In the well known classification of compact 2-manifolds, any such manifold is homeomorphic to a sphere with g handles (if orientable) or to a sphere with g cross-caps (if non-orientable)*. Hence, the Euler characteristic χ for a closed surface (i.e. a surface without a boundary) is also given by

$$\begin{aligned} \chi &= 2(1 - g) && \text{(for an orientable surface)} \\ &= \quad 2 - g && \text{(for a non-orientable surface).} \end{aligned} \tag{3}$$

The cases compatible with non-negative integral solutions for χ are thus exactly four in number. The only surfaces admitting finite fullerene maps in the sense of our definition are therefore: S^2 (the sphere, orientable with $g = 0$), T^2 (the torus, orientable with $g = 1$), K^2 (the Klein bottle, non-orientable with $g = 2$) and P^2 (the real projective plane, also called the elliptic plane, non-orientable with $g = 1$). All embeddings are 2-cell, meaning that each face is homeomorphic to an open disk. An immediate consequence of Euler's formula is that fullerenes on S^2, T^2, K^2 and P^2 have exactly 12, 0, 0 and 6 pentagons, respectively.

---

* Handles are made from cylinders and cross-caps from twisted eylinders (see ref. [9] for details).

The four possible classes of fullerenes are therefore *spherical, toroidal, Klein-bottle* and *elliptic*. No other surfaces are compatible with the definition. Toroidal and Klein-bottle fullerenes may also be called toroidal and Klein-bottle *polyhexes* [10,11] as they include no pentagons.

Maps on S^2 can be drawn as the usual Schlegel diagrams, and maps on T^2, K^2 and P^2 by identifying opposite edges of a fundamental parallelogram with appropriate orientation. Maps on P^2 are more usually drawn inside a circular frame where antipodal boundary points are to be identified. Fig. 1 shows examples of small fullerenes from the four classes, drawn as the graph, the map and its dual triangulation in the appropriate surface. We remark that the Petersen and Heawood graphs which appear naturally here are actually the 5- and 6-cages (a k-cage is a trivalent graph of smallest cycle size k with the largest possible number of edges); their duals in P^2 and T^2, K_6 and K_7, realise the chromatic number of the corresponding surfaces.

Spherical and toroidal fullerenes have an extensive chemical literature, and Klein-bottle polyhexes have been considered in several papers [11-13]. The review chapter by Klein and Zhu [13] in particular, introduces many of the relevant concepts from surface topology to a chemical context. Elliptic fullerenes have appeared so far only in ref. [8], but turn out to be related in a simple way to a subclass of the known spherical fullerenes, as shown later.

Spherical fullerenes

It has been proved that at least one spherical fullerene with n vertices (modelling a carbon molecule C_n) exists for all even n with $n \geq 20$ except for the case $n = 22$ [14]. Each fullerene polyhedron has $f_5 = 12$ pentagons and $f_6 = n/2 - 10$ hexagons. Chemical interest centres on isolated-pentagon fullerenes, which can be constructed for $n = 60$ and for all even values of $n \geq 70$ (thus with $f_6 = 20$ and $f_6 \geq 25$). Aspects of the systematics of spherical fullerenes including chemical results are summarised in, e.g. ref. [2].

Toroidal and Klein-bottle fullerenes

T^2- and K^2-polyhexes are related to the hexagonal tessellation of the graphite sheet in a straightforward way. The underlying surfaces are quotients of the Euclidean plane R^2 under groups of isometries generated by two translations (for T^2) or one translation and one glide reflection (for K^2). Each point of T^2 and K^2 corresponds to an orbit of the generating group. For completeness, we note that the groups generated by a single translation or a single glide reflection respectively give as quotients the cylinder and the twisted cylinder (the Möbius surface). Construction and enumeration of polyhexes can

therefore be envisaged as a process of cutting parallelograms out of the graphite plane and gluing their edges according to the rules implied in Fig. 1.

Some confusion exists in the mathematical and chemical literature of toroidal polyhexes. Negami [15], Altschuler [16] and other topological graph theorists define regular 3-valent maps on the torus to mean 2-cell embeddings with all faces hexagonal, without further qualification. Errera [17], Brahana [18], Coxeter [19] and others working in a group theoretical tradition use the same term in a more restricted sense of polyhexes with automorphism groups G of the maximal possible order $|Aut(G)|$, in other words, those that realise the equality in the analogue of the Weinberg bound $|Aut(G)| \leq 4e(G)$ ($= 6f_6$ for a polyhex). These regular maps are called *fully symmetric* by Nakamoto in his thesis [20]. All such fully symmetric graph embeddings are: (on S^2) the five Platonic polyhedra, (on P^2) six graphs that include the Petersen graph and its dual, (on K^2) no graphs at all [20], and (on T^2) the polyhexes that arise from an analogue of the Goldberg/Coxeter construction of icosahedral S^2-fullerene polyhedra [21,22].

Here we consider only *polyhedral* polyhexes, i.e. those without loops or multiple edges and where the intersection of any two faces is either one edge or is empty. The dual of a toroidal polyhedral polyhex is a triangulation of the torus. To illustrate the relationship of the various definitions, we give the counts for small cases in Table 1, using data extracted from the papers of Negami [15] and Altschuler [16]. The tabulations given by Kirby [10,12] include some non-polyhedral cases.

In Negami's construction, a three-parameter code [15] represents any toroidal polyhex (or, equivalently, any 6-regular triangulation of T^2) as a tessellation of the hexagonal lattice. Each graph of this type is denoted $T(p,q,r)$, with integer parameters p, q and r where p is the length of a geodesic cycle of edge-sharing hexagons, r is the number of such cycles and q is an offset.

Polyhex maps on T^2 are constructible for all values $f_6 \geq 3$, $n \geq 6$ [23]. At least one *polyhedral* toroidal polyhex exists for all even numbers of vertices $n \geq 14$. The unique polyhedral toroidal fullerene at $n = 14$ is a realisation of the Heawood graph. It has indices $(2,1)$ in the Goldberg/Coxeter construction and is the dual of K_7, the complete graph on seven vertices, which itself realises the 7-colour map on the torus. This map and its dual are shown in Fig. 1. A different presentation of this and the next three toroidal polyhexes obtainable by the Goldberg/Coxeter construction are illustrated in Fig. 2.

A description of Klein-bottle polyhexes can be developed along similar lines [20,24]. Each toroidal graph $T(p,0,r)$ can be used to obtain two Klein-bottle 6-regular triangula-

4

tions (and hence, by dualisation, fullerenes), the *handle* and *cross-cap* types $K_h(p,r)$ and $K_c(p,r)$, respectively. The torus is cut along a geodesic of length p. Then the *handle* construction amounts to identification of opposite sides of the resulting parallelogram with reversed direction. In the *cross-cap* construction, the opposite sides are each converted to cross caps, with slightly different rules for odd and even p. See also ref. [25] for pictures of the two types of Klein bottle. Polyhex maps on K^2 are constructible for all values $f_6 \geq 3$, $n \geq 6$ [23]. The unique smallest *polyhedral* Klein-bottle polyhex has 18 vertices (9 hexagonal faces) and is the dual of the tripartite $K_{3,3,3}$; the graph, the map and its dual are shown in Fig. 1.

It will turn out to be useful for calculation of spectra later that each Klein-bottle polyhex graph, whether of handle or cross-cap type, has a double cover among the centrosymmetric toroidal polyhexes. In contrast with a T^2-polyhex, a polyhex on K^2 may or may not be bipartite, i.e. spanned by two disjoint sets of vertices, black and white, such that every white vertex is surrounded by black and *vice versa*.

Elliptic fullerenes

Torus and Klein-bottle arise as quotient spaces, as described above, and this leads to a construction of the possible polyhex maps. Within the same framework the real projective plane arises as a quotient space of the sphere, the required group being C_i. The real projective plane (also known as the elliptic plane) is obtained by identifying antipodal points of the spherical surface; in other words, it is the *antipodal quotient* of the sphere. P^2 is the simplest compact non-orientable surface in the sense that it can be obtained from the sphere by adding just one cross-cap.

Clearly, this construction can be carried over to maps: the antipodal quotient of a centrosymmetric map on the sphere has vertices, edges and faces obtained by identifying antipodal vertices, edges and faces, thereby halving the number of each type of structural component. For example, the antipodal quotient of the icosahedron is K_6, the complete graph on 6 vertices, and that of the dodecahedron is the Petersen graph, famous as a counterexample to many theorems. The Petersen graph is not planar but it is called *projective-planar* in the sense that it can be embedded without edge crossings in the real projective plane.

In this terminology, our definition of elliptic fullerenes amounts to selection of polyhedral projective-planar trivalent maps with only 5- and 6-gonal faces. As noted above, $f_5 = 6$ for these maps. Thus, the Petersen graph is an elliptic fullerene (the smallest). Maps on P^2 with $f_5 = 6$ are constructible for $f_6 = 0$ and for all values $f_6 \geq 3$ [23]. Not

5

all of these are polyhedral. In general the elliptic fullerenes are exactly the antipodal quotients of the centrosymmetric spherical fullerenes. From any centrosymmetric fullerene it is possible to construct a unique elliptic fullerene by identifying antipodal vertices, and from any elliptic fullerene it is possible to reconstruct uniquely the original centrosymmetric fullerene.

Thus, the problem of enumeration and construction of elliptic fullerenes reduces simply to that for centrosymmetric conventional spherical fullerenes. The point symmetry groups that contain the inversion operation are C_i, C_{nh}, (n even), D_{nh} (n even), D_{nd} (n odd), T_h, O_h and I_h. A spherical fullerene may belong to one of 28 point groups [2] of which 8 appear in the previous list: C_i, C_{2h}, D_{2h}, D_{6h}, D_{3d}, D_{5d}, T_h and I_h. Clearly, a fullerene C_n can be centrosymmetric only if n is divisible by 4 as f_6 must be even but also $f_6 = n/2 - 10$. After the minimal case $n = 20$, it turns out that there are no centrosymmetric fullerenes at $n = 24$ and $n = 28$, and unique examples at 32 (D_{3d}) and 36 (D_{6h}). Complete enumerations of general centrosymmetric fullerenes on up to 100 atoms and of isolated pentagon centrosymmetric fullerenes of up to 140 atoms taken from the Fullerene Atlas [2] are given in Tables 2 and 3. All generate elliptic fullerenes by antipodal identification.

It seems likely that at least one centrosymmetric fullerene exists for all doubly even values of $n \geq 32$ and at least one centrosymmetric fullerene with isolated pentagons for every doubly even $n \geq 92$, though we are not aware of a proof. For four of the eight point groups, explicit conditions are known for the existence of a fullerene of the given symmetry at a given n: I_h fullerenes C_n exist for all distinct solutions (i, j) of $n = 20(i^2 + ij + j^2)$ with either $i = j$ or $j = 0$, and similar but more complicated conditions are known for T_h, D_{5d} and D_{6h} fullerenes [26].

Some infinite analogues of fullerenes

All the fullerenes considered so far are finite and are actually trivalent tilings with (combinatorial) pentagons and hexagons of particular surfaces. If instead of S^2, T^2, K^2 or P^2, we consider tilings of the Euclidean plane R^2, a natural definition is obtained for an *infinite* fullerene analogue. Namely, a *plane fullerene* is a trivalent tiling of R^2 by combinatorial pentagons and hexagons. Deza and Shtogrin [8] proved that the number of pentagons in a planar fullerene is at least 6. This follows from an old result of A.D. Alexandrov [27]. It is easy to see that the plane fullerenes with $f_5 = 0$ (the graphite sheet), and $f_5 = 1$ (a pentagonal cone) are unique. However, there is an infinity of possibilities for $2 \leq f_5 \leq 6$. Restriction to bounded tile size eliminates pathological possibilities such as an infinite tube

capped by a hemi-dodecahedron. The restriction to trivalence is also a powerful one: by allowing also four-valent vertices, for example, R^2 could be partitioned into pentagons in a tiling with vertices of degrees three and four (Fig. 3) which would not be a fullerene in our sense.

Other infinite fullerenes would be given by trivalent tilings with pentagons and hexagons of the cylinder, semi-infinite cylinder, twisted cylinder etc. Such a tiling on the cylinder is a polyhex and is the infinite open nanotube.

The tiling description also leads naturally to definition of a *space fullerene* as a four-valent tiling of R^3 where each cell is a conventional fullerene polyhedron. It turns out that those space fullerenes where the cells have no adjacent *hexagons* [C_{20}, C_{24}, C_{26}, $C_{28}(T_d)$] are of special interest, though others have been constructed [8]. These space fullerenes occur in chemistry and physics as the 'dodecahedral family of hydrates' [28] (clathrates) and their duals as 'tetrahedrally close-packed phases' (t.c.p) or generalised Frank-Kasper phases [29-31]. If the inventory of cells is extended to the C_{22} 'near fullerene' (the edge-truncated dodecahedron with 1 square, 10 pentagonal and two hexagonal faces), other phases can be represented, e.g. Hume-Rothery's phase γ has $C_{20} : C_{22} : C_{26}$ in the proportion $2 : 2 : 3$ [31].

In the hydrate structures the 'vertices' are water molecules with two donor and two acceptor hydrogen bonds. Each space fullerene gives rise to a hypothetical silicate structure if every vertex is replaced by an SiO_4 tetrahedron, or a hypothetical carbon allotrope if every vertex is replaced by a single carbon atom. Combinations tabulated by Wells [28] can be represented in an obvious 'chemical' fullerene notation as $(C_{20})(C_{24})_3$, $(C_{20})_2(T_d\ C_{28})$, $(C_{20})_3(C_{24})_2(C_{26})_2$ and $(C_{20})_5(C_{24})_8(C_{26})_2$. The first of these is illustrated in Fig. 4. A new structure consisting of C_{20}-, C_{24}- and $C_{36}(D_{6h})$- cells is given by Deza and Shtogrin [8]. Four-valent honeycombs with fullerene and similar cells also figure in modern conjectured solutions to the Kelvin problem of finding a partition of three-dimensional space into cells of equal volume and minimal surface area (see also other papers in the volume containing refs. [30] and [32]). The best foam found so far [32] is the dual of the A15 Frank-Kasper structure and is a metric variation of $(C_{20})(C_{24})_3$.

Generalising further, the vertices could be replaced by larger entities such as tetra-hedral C_{28} fullerenes bonded through their four apical atoms to make a super-fullerene lattice. The 3D tilings open up a number of questions of enumeration, characterisation and spectral structure, to answer which will require further work.

We note that plane fullerenes can be seen as infinite fullerene polyhedra, and space

7

fullerenes as an infinite analogue of a four-dimensional fullerene. A four-dimensional fullerene (a *polytopal 4-fullerene* in the language of [8]) is therefore a simple 4-polytope having only five- and six-sided two-faces. Clearly, all cells of such structures are fullerenes.

Eigenvalue properties

A first indication of the qualitative π-electronic structure to be expected of the new frameworks as hypothetical forms of carbon can be gained from Hückel theory, for which a pre-requisite is a knowledge of the adjacency properties. For orientable surfaces there is a clear link between the spectrum of the adjacency matrix of the map and the π-orbital energies of its realisation as a carbon framework. In the simplest Hückel model, each eigenvalue λ of the matrix corresponds to an orbital of energy $\alpha + \lambda\beta$, where α is the Coulomb parameter, assumed the same for every site, and β the resonance parameter, assumed equal for all bonds. For non-orientable surfaces such as K^2 and P^2 this correspondence is lost because the intrinsic twist in the surface introduces a phase discontinuity in the π-basis. The relevant eigenvalues are then those of a weighted adjacency matrix, as will be discussed below.

It should be remembered that π energy is only one contribution to the total energy, and in real spherical fullerenes it is dominated by the strain in the σ system, which leads for example to the observation that stable fullerenes are not necessarily those of maximal Hückel energy [2]. As real carbon systems, T^2-fullerenes are highly strained and are unlikely to be realised unless for very large values of n; the toroidal nanotube reported by Liu et al. [7] has a diameter of $330 - 500$ nm, implying many thousands of atoms. Chemical systems based on K^2- and P^2-polyhexes are less plausible, as these systems involve self-intersection. Klein [11] suggests a mode of interlocking of graphitic planes that may minimise the very considerable energetic costs, but the interest of these systems is likely to remain purely mathematical for a long while to come.

(a) Orientable fullerenes

S^2-*fullerenes.* The situation for the usual spherical fullerenes has been well explored. Adjacency matrices of spherical fullerenes have typically more positive than negative eigenvalues, correlating with their chemical behaviour as electron-deficient π-systems. Only occasional examples with more negative than positive eigenvalues are known [33]. A special subclass with equal numbers of positive and negative eigenvalues, and therefore an 'ideal' π-structure, is formed by the *leapfrog* fullerenes C_n, each constructed by omnicapping and then dualising a smaller S^2-fullerene $C_{n/3}$ [34]. Other spherical fullerenes with exactly $n/2$ positive eigenvalues are possible, but are rare compared to the leapfrogs [2].

8

Leapfrogging can be carried out on any surface, with characteristic implications for the eigenvalue spectrum. As an illustration of the leapfrog construction on non-spherical surfaces, the leapfrogs of the smallest spherical, toroidal, Klein-bottle and elliptic fullerenes are given in Fig. 5. The striking spectral regularities from the leapfrog transformation can be rationalised in terms of the way that relationships between structural components in a parent carry over to the leapfrog map. Each face of the parent gives rise to a congruent but rotated face in the leapfrog; these *Clar* faces are disjoint and exhaust the vertices of the leapfrog. All faces of the leapfrog outside the Clar set are hexagons centred on the sites of the parent vertices. Each edge of the parent gives rise to a rotated edge in the leapfrog; these *Fries* edges are again disjoint, and account for one third of the edges of the leapfrog and all of its vertices. The Fries edges radiate from the Clar faces, so that every edge of the leapfrog is either Fries or *Clar* (i.e. an edge of a Clar face).

A consistent Kekulé structure can be built for leapfrog fullerenes on any of the four surfaces by placing formal double bonds on the Fries edges and formal single bonds on the Clar edges. This *Fries* structure has the maximum possible number of simultaneous benzenoid hexagons, one for each vertex of the parent, giving an ideal localised electronic structure for the neutral carbon cage C_n. A delocalised version of the argument uses considerations based on the Rayleigh inequality for the distinct basis sets consisting of all bonding (in-phase) or all anti-bonding (out-of-phase) combinations along Fries edges [35] and shows that S^2-leapfrog fullerenes have *no* zeros and hence closed shells as neutral molecules.

A second localised structure places a sextet of π electrons on every Clar face and a single bond on every Fries edge; this is a formal model of the electronic structure of the anionic system $C_n^{f_5-}$ bearing an excess of f_5 electrons. This too has its counterpart in delocalised molecular-orbital theory, where the extra 12 electrons of a leapfrog S^2-fullerene anion C_n^{12-} occupy low-lying anti-bonding orbitals of translational and rotational symmetry [36].

T^2-fullerenes. Eigenvalue spectra of toroidal polyhexes have been studied in some detail. Kirby et al. [10] give an explicit formula for calculation of the set of eigenvalues in terms of canonical lattice-vector parameters. T^2-polyhexes have symmetric spectra (with both $+\lambda$ and $-\lambda$ occurring for every eigenvalue λ), as they are bipartite graphs, and all eigenvalues λ except ± 3 and ± 1 have even multiplicity. The special eigenvalue $\lambda = 0$ is governed by a simple pattern: exactly those toroidal polyhedral polyhexes that are leapfrogs have open-shells, with four zero eigenvalues at positions $n/2 - 1$, $n/2$, $n/2 + 1$ and $n/2 + 2$ in the spectrum [37]. The spectra of toroidal polyhexes are also intimately

9

related to those of spherical triangle and hexagon polyhedra, and explain endospectral regularities in the latter series [38].

(b) Non-orientable fullerenes

As warned earlier, there is a subtlety in the application of Hückel theory to frameworks embedded in non-orientable surfaces. The problem arises as follows. The π-basis consists of a p function on every participating carbon atom, directed along the normal to the surface, and Hückel energies are obtained by diagonalising a Hamiltonian matrix whose entries are pairwise integrals of these (vector-like) functions. Under the usual assumptions, only functions on nearest neighbours are involved. For an orientable surface, neglecting curvature, neighbouring normals are parallel and the integrals are therefore proportional to entries in the adjacency matrix. However, for a non-orientable surface made by gluing edges of a patch together with a twist, p functions neighbouring across the join become antiparallel and for such pairs the integral is reversed in sign (see Fig. 6).

A standard chemical example occurs in the theory of Möbius transition states for pericyclic processes where eigenvalues for cycles with one phase interuption are found by diagonalising a weighted adjacency matrix that has an entry -1 for one link and $+1$ for all others. The spectrum of the Möbius cycle can be found by rotation of the usual geometric construction for untwisted cycles [39]. In the present case, an analogous procedure can be adopted: to construct the dimensionless Hückel Hamiltonian matrix, $\mathbf{H}$, take the adjacency matrix $\mathbf{A}$ of the graph and multiply by -1 all entries for edges that cross a twisted boundary (in the case of K^2), or cross the circular boundary (in the case of P^2). Signs for any edges terminating at or lying within a boundary can be decided by making small shifts to bring their vertices inside the boundary.

K^2-fullerenes. Some calculations of eigenvalues of unweighted adjacency matrices of Klein-bottle polyhexes have been reported [40] and compared with those of toroidal polyhexes, but a general picture of the spectra for these systems has not been given.

The Klein-bottle surface can be obtained by identifying diametrically opposite points of the torus, i.e. by collapsing each point and its antipode [41]. The point groups available to the covering torus are at most the centrosymmetric subgroups of $D_{\infty h}$ i.e. D_{nh} and C_{nh} (n even), D_{nd} and S_{2n} (n odd), C_i, though, as with spherical fullerenes, some of the lower groups may not be realisable. Hence, each K^2-polyhex on n vertices is doubly covered by a centrosymmetric T^2-polyhex on $2n$ vertices, and is therefore a *divisor* [42] of the larger graph. By a centrosymmetric graph, we mean, as usual, a graph that has a centrosymmetric setting on the appropriate surface, i.e. has centrosymmetric maximal symmetry. The

10

weighted K^2-polyhex represented by the $\mathbf{H}$ matrix is the co-divisor. Any given eigenvector of the adjacency matrix of the larger T^2-polyhex therefore corresponds to an eigenvalue in the spectrum of $\mathbf{A}$ ($\mathbf{H}$) of the smaller K^2-graph if the antipodal vertices carry equal (opposite) coefficients. Eigenvectors of the covering graph can always be projected to display this *gerade/ungerade* symmetry. Fig. 7 shows the construction of double covers for two small Klein-bottle polyhexes. Given a K^2-polyhex, the double cover can always be constructed, reduced to canonical form and its spectrum partitioned into that of $\mathbf{A}$ (K^2) and $\mathbf{H}$ (K^2).

A simple pattern can be observed in the partitioned spectra. Consider a given eigenvector $|\lambda >$ with eigenvalue λ in the toroidal double cover. As the covering T^2-polyhex is bipartite, a change of sign of the coefficient on every vertex of one partite set generates an eigenvector $| - \lambda >$ of eigenvalue $-\lambda$. On collapsing pairs of covering vertices, $| + \lambda >$ will generate an eigenvector of either $\mathbf{A}$ or $\mathbf{H}$, as will $| - \lambda >$. Two different situations are possible:

(i) *the K^2-polyhex itself is bipartite*: $|+\lambda >$ and $|-\lambda >$ yield eigenvectors with eigenvalues belonging to one and the same subspectrum, either the spectrum of $\mathbf{A}$ or the spectrum of $\mathbf{H}$. For example, in this case, the $\mathbf{A}$ spectrum always contains $\lambda = 3$ and $\lambda = -3$, whereas the $\mathbf{H}$ spectrum contains neither.

(ii) *the K^2-polyhex itself is non-bipartite*: $| + \lambda >$ and $| - \lambda >$ yield eigenvectors with eigenvalues belonging to different subspectra, one to the spectrum of $\mathbf{A}$ and one to that of $\mathbf{H}$. As a consequence, the two subspectra are exact reversals of one another. In particular, the $\mathbf{A}$ spectrum always contains $\lambda = +3$ and the $\mathbf{H}$ spectrum $\lambda = -3$.

The origin of the properties (i) and (ii) is readily explained. Take the patch of hexagons that generates the K^2-polyhex, and colour its vertices alternately black and white. Further, take an eigenvector $|\lambda >$ of the covering torus that yields a vector of the same eigenvalue for $\mathbf{A}$. $|\lambda >$ will have the property that any one pair of antipodal vertices on the torus share a coefficient. Now attach to each vertex of the patch the common value of the coefficient from its covering pair, to produce a self-consistent eigenvector of $\mathbf{A}$, $|\lambda(\mathbf{A}) >$. If the K^2-polyhex is bipartite (case (i)), reversal of the coefficients of all black vertices of the patch gives another self-consistent eigenvector of $\mathbf{A}$, $| - \lambda(\mathbf{A}) >$. However, if the K^2-polyhex is non-bipartite (case (ii)), then when the patch is joined up to make the Klein bottle each edge of the graph that crosses the twisted boundary will join vertices of like colour, since it is these edges that destroy the alternating pattern of the planar patch. In such a case, weighting these edges by -1 and simultaneously reversing all coefficients

11

on the vertices of one colour gives an eigenvector of $\mathbf{H}$ with eigenvalue $-\lambda$, i.e. $|-\lambda(\mathbf{H}) >$. QED.

A leapfrog rule can be derived for K^2-polyhexes. Consider a patch cut from the hexagonal tessellation of the plane, such that it can be rolled up to give both leapfrog T^2- and K^2-polyhexes. This possibility implies that the vertices of the patch are spanned by Clar hexagons. Fig. 8 shows such a patch. When wrapped as a torus, the polyhex must have four zero-eigenvalue vectors, as it is a leapfrog. Fig. 8 (c)−(f) shows their explicit form. In terms of the Fries edges, both bonding and anti-bonding spaces each contribute two adjacency eigenvectors of zero eigenvalue. Inspection of the figure shows that different subsets of exactly two of the four survive as eigenvectors of either $\mathbf{A}$ or $\mathbf{H}$ matrices when the same polyhex is glued as a K^2-fullerene. The Rayleigh inequality arguments used in ref. [35,43] show that, as bonding and anti-bonding spaces each contribute one zero in the Klein-bottle form, the two zeroes are eigenvalues $n/2$ and $n/2+1$, i.e. HOMO and LUMO of the hypothetical neutral carbon framework with this topology. The general chemical conclusion is that *all leapfrog K^2-polyhexes, whether treated as weighted or unweighted Hückel problems, have open shells, with two electrons in two non-bonding π orbitals.*

A final feature of the leapfrog transformation is illustrated by Fig. 9, where the graphs of the 18-vertex K^2-fullerene, its leapfrog and double leapfrog are superimposed. Leapfrogging switches the character of the graph from bipartite to non-bipartite and back again. This is a result of a switch in parity of geodesic cycles, even though all faces remain hexagonal, and is part of a more general pattern: leapfrogging a cubic graph with all faces even switches the bipartite character on the non-orientable surfaces (K^2 and P^2) but leaves it unchanged on the orientable surfaces (S^2 and T^2).

P^2-fullerenes. The eigenvalue spectra, both weighted and unweighted, of an elliptic fullerene are immediately available from the adjacency spectrum of its centrosymmetric parent. The eigenvalues of $\mathbf{A}$ for a P^2-fullerene are just those of the parent that correspond to *gerade* eigenvectors; eigenvalues of $\mathbf{H}$ correspond to *ungerade* eigenvectors of the parent. Together, the weighted and unweighted spectra of the P^2-fullerene sum to the spectrum of the parent, since the smaller graph is a divisor of the larger.

A simple consequence is that any P^2-fullerene derived from a leapfrog spherical parent has a properly closed shell as a neutral π system. Proof: the leapfrog parent C_n has $n/2$ positive and $n/2$ negative eigenvalues [35]. Its bonding eigenvectors span the permutation representation of the Fries edges [34,36]. This representation has character zero under inversion as all edges shift under this operation. Hence, the centrosymmetric leapfrog S^2-

fullerene C_n has $n/4$ gerade and $n/4$ ungerade bonding eigenvectors. Therefore, the spectra of $\mathbf{A}$ and $\mathbf{H}$ of the derived P^2-fullerene will each have $n/4$ bonding, $n/4$ anti-bonding and zero non-bonding eigenvectors, QED. As the operation that derives P^2- from S^2-fullerenes commutes with the leapfrog transformation, this argument proves that *leapfrog P^2-fullerenes have properly closed shells.*

Conclusion

This paper has explored the extension of the fullerene concept from the sphere to other surfaces, retaining trivalence and the limitation to face sizes five and six, showing that the chemical species exist within a mathematical context of a limited set of possibilities for tiling. Consideration of the extended set of surfaces also gives a context to the magic-number rules of Hückel theory, such as the leapfrog rule for closed-shell S^2-fullerenes, as it turns out that leapfrog polyhedra have distinct but predictable properties on all four surfaces. Extensions to other face sizes and more exotic surfaces will allow description of many more variants on the sphere, some of which are promising as candidates for carbon polyhedra or infinite solids.

Acknowledgment

References

1. T. Braun, A. Schubert, H. Maczelka and L. Vasvári, Fullerene research 1985-1993, World Scientific, Singapore, 1995.

2. P.W. Fowler and D.E. Manolopoulos, An atlas of fullerenes, Oxford University Press, Oxford, 1995.

3. R. Taylor, ed., The chemistry of fullerenes, World Scientific, Singapore, 1995.

4. M.S. Dresselhaus, G. Dresselhaus and P.C. Eklund, Science of fullerenes and carbon nanotubes, Academic Press, San Diego, 1996.

5. P.W. Fowler, D.E. Manolopoulos, G. Orlandi and F. Zerbetto, Energetics and isomerisation pathways of a lower fullerene: the Stone-Wales map for C_{40}, J. Chem. Soc. Faraday Trans. 91 (1995) 1421-1423.

6. A. Ayuela, P.W. Fowler, D. Mitchell, R. Schmidt, G. Seifert and F. Zerbetto, C_{62}: theoretical evidence for a non-classical fullerene with a heptagonal ring, J. Phys. Chem. 100 (1996) 15634-15636.

7. J. Liu, H. Dai, J.H. Hafner, D.T. Colbert, R.E. Smalley, S.J. Tans and C. Dekker, Fullerene crop circles, Nature 385 (1997) 780-781.

8. M. Deza and M.I. Shtogrin, Three-, four- and five-dimensional fullerenes, SE Asian Bull. Math. 23 (1999) 1-10.

9. J.L. Gross and T.N. Tucker, Topological graph theory, Wiley, New York, 1987.

10. E.C. Kirby, R.B. Mallion and P. Pollak, Toroidal polyhexes, J. Chem. Soc. Faraday Trans. 89 (1993) 1945-1953.

11. D.J. Klein, Elemental benzenoids, J. Chem. Inf. Comp. Sci. 34 (1994) 453-459.

12. E.C. Kirby, Recent works on toroidal and other exotic fullerene structures, in: From chemical topology to 3-dimensional geometry, A.T. Balaban, ed., Plenum Press, New York, 1997, Chapter 8, 263-296.

13. D.J. Klein and H. Zhu, All-conjugated carbon species, in: From chemical topology to 3-dimensional geometry, A.T. Balaban, ed., Plenum Press, New York, 1997, Chapter 9, 297-341.

14. B. Grünbaum and T.G. Motzkin, The number of hexagons in and the simplicity of geodesics on certain polyhedra, Can. J. Math. 15 (1963) 744-751.

15. S. Negami, Uniqueness and faithfulness of embedding of toroidal graphs, Discrete Math. 44 (1983) 161-180.

16. A. Altshuler, Construction and enumeration of regular maps on the torus, Discrete Math. 4 (1973) 201-217.

17. A. Errera, Sur les polyhedres reguliers de l'analysis situs, Acad. Roy. Belg. Cl. Sci. Mem. Coll. 8(2) 7 (1922) 1-17.

18. H.R. Brahana, Regular maps on the anchor ring, Amer. J. Math. 48 (1926) 225-240.

19. H.S.M. Coxeter and W.O.J. Moser, Generators and relations for discrete groups, 2nd edition, Springer, Berlin, 1965.

20. A. Nakamoto, Triangulations and quadrangulations of surfaces, D.Sc. thesis, Dept. of Mathematics, Keio University, Japan, 1996.

21. M. Goldberg, A class of multi-symmetric polyhedra, Tôhoku Math. J. 43 (1937) 104-108.

22. H.S.M. Coxeter, Virus macromolecules and geodesic domes, in: A spectrum of mathematics, J.C. Butcher, ed., Oxford University Press/Auckland University Press, Ox-

ford/Auckland, 1971, 98-107.

23. F. Plastria, On the number of hexagons in cubic maps, Report BEIF/98, Centrum voor Bedrijfsinformatie, Vrije Universiteit Brussel, 1998.

24. S. Negami, Classification of 6-regular Klein bottle graphs, Res. Rep. Inf. Sci. Tokyo Institute of Technology A96 (1984).

25. J. Stillwell, Classical topology and combinatorial group theory, 2nd edition, Springer, Berlin, 1993 (see pp. 65-67).

26. P.W. Fowler, J.E. Cremona and J.I. Steer, Systematics of bonding in non-icosahedral carbon clusters, Theo. Chim. Acta 73 (1988) 1-26.

27. See, for example: A.D. Alexandrov, Convex Polyheder, Akademie-Verlag, Berlin, 1958, p. 92.

28. A.F. Wells, Structural Inorganic Chemistry, 4th edition, Oxford University Press, Oxford, 1975.

29. D.P. Shoemaker and C.B. Shoemaker, Concerning the relative numbers of atomic coordination types in tetrahedrally close-packed metal structures, Acta Cryst. B42 (1986) 3-11.

30. N. Rivier and T. Aste, Organised packing, in: The Kelvin problem, D. Weaire, ed., Taylor & Francis, London, 1996, 61-68 (see especially Table 1).

31. J.F. Sadoc and R. Mosseri, Frustation géométrique, Eyrolles, Paris, 1997 (see Chapter 7, especially Table 1).

32. R. Kusner and J.M. Sullivan, Comparing the Weaire-Phelan equal-volume foam to Kelvin's foam, in: The Kelvin problem, D. Weaire, ed., Taylor & Francis, London, 1996, 71-80.

33. P.W. Fowler, Fullerene graphs with more negative than positive eigenvalues: the exceptions that prove the rule of electron deficiency?, J. Chem. Soc. Faraday Trans. 93 (1997) 1-3.

34. P.W. Fowler and J.I. Steer, The leapfrog principle: a rule for electron counts of carbon clusters, J. Chem. Soc. Chem. Comm. (1987) 1403-1405.

35. D.E. Manolopoulos, D.R. Woodall and P.W. Fowler, Electronic stability of fullerenes: eigenvalue theorems for leapfrog carbon clusters, J. Chem. Soc. Faraday Trans. 88 (1992) 2427-2435.

36. P.W. Fowler and A. Ceulemans, Electron deficiency of the fullerenes, J. Phys. Chem. 99 (1995) 508-510.

37. M. Yoshida, M. Fujita, P.W. Fowler and E.C. Kirby, Non-bonding orbitals in graphite, carbon tubules, toroids and fullerenes, J. Chem. Soc. Faraday Trans. 93 (1997) 1037-1043.

38. P.W. Fowler, P.E. John and H. Sachs, $(3,6)$ cages, hexagonal toroidal cages and their spectra, in: Discrete Mathematical Chemistry, P. Hansen, P.W. Fowler and M. Zheng, eds., DIMACS Series on Discrete Mathematics and Theoretical Computer Science, AMS, 1999.

39. For a simple geometric construction of the spectrum, see for example: G.B. Gill and M.R. Willis, Pericyclic reactions, Chapman and Hall, London, 1974.

40. E.C. Kirby, Remarks upon recognising genus and possible shapes of chemical cages in the form of polyhedra, tori and Klein bottles, Croat. Chem. Acta 68 (1995) 269-282.

41. D. Hilbert, S. Cohn-Vossen, Geometry and the imagination, 2nd edition, Chelsea Publishing Co., New York, 1990 (see Fig. 300, p. 312).

42. D.M. Cvetković, M. Doob and H. Sachs, Spectra of graphs: theory and application, Academic Press, New York, 1979.

43. P.W. Fowler and K.M. Rogers, Eigenvalue spectra of leapfrog polyhedra, J. Chem. Soc. Faraday Trans. 94 (1998) 2509-2514.

Table 1: Enumeration of toroidal polyhexes. f_6 is the number of hexagonal faces. The count of regular polyhexes is for 2-cell embeddings of trivalent maps with all faces hexagonal. Polyhedral polyhexes are those with dual triangulations in which the intersection of any two faces is either a vertex or is empty. The final rows give the counts and canonical lattice-vector parameters for the restricted class of regular polyhexes of maximal automorphism group.

f_6	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Regular	1	1	2	3	2	3	3	5	4	4	3	8	4	5
Polyhedral	-	-	-	-	-	-	1	1	2	1	1	4	2	2
Fully symmetric	1	-	1	1	-	-	1	-	1	-	-	1	1	-
	$(1,0)$	-	$(1,1)$	$(2,0)$	-	-	$(2,1)$	-	$(3,0)$	-	-	$(2,2)$	$(3,1)$	-

Table 2: Centrosymmetric fullerenes C_n ($20 \leq n \leq 100$) with and without pentagon adjacencies. Each is the parent of an elliptic P^2-fullerene.

n	C_i	C_{2h}	D_{2h}	D_{6h}	D_{3d}	D_{5d}	T_h	I_h	Total
20	0	0	0	0	0	0	0	1	1
24	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0
32	0	0	0	0	1	0	0	0	1
36	0	0	0	1	0	0	0	0	1
40	0	0	1	0	0	2	0	0	3
44	0	0	0	0	3	0	0	0	3
48	0	1	2	0	0	0	0	0	3
52	0	1	2	0	0	0	0	0	3
56	1	2	1	0	2	0	0	0	6
60	0	4	1	2	0	1	0	1	9
64	2	4	1	0	0	0	0	0	7
68	0	7	2	0	3	0	0	0	12
72	3	7	5	0	0	0	0	0	15
76	2	11	0	0	0	0	0	0	13
80	5	16	0	0	2	2	0	1	26
84	9	4	6	2	1	0	0	0	22
88	10	16	5	0	0	0	0	0	31
92	12	13	4	0	5	0	1	0	35
96	20	16	3	2	1	0	0	0	42
100	14	28	2	0	0	2	0	0	46
Total	78	130	35	7	18	7	1	3	279

Table 3: Centrosymmetric fullerenes C_n $(60 \leq n \leq 140)$ with isolated pentagons. Each is the parent of an elliptic P^2-fullerene.

n	C_i	C_{2h}	D_{2h}	D_{6h}	D_{3d}	D_{5d}	T_h	I_h	Total
60	0	0	0	0	0	0	0	1	1
72	0	0	0	0	0	0	0	0	0
76	0	0	0	0	0	0	0	0	0
80	0	0	0	0	0	1	0	1	2
84	0	0	0	1	1	0	0	0	2
88	0	0	0	0	0	0	0	0	0
92	0	0	1	0	0	0	0	0	1
96	0	0	1	1	1	0	0	0	3
100	0	0	0	0	0	1	0	0	1
104	0	0	2	0	1	0	0	0	3
108	0	3	1	1	1	0	0	0	6
112	0	2	1	0	0	0	0	0	3
116	0	2	1	0	0	0	1	0	4
120	1	5	3	0	0	3	0	0	12
124	1	5	1	0	0	0	0	0	7
128	3	4	0	0	2	0	0	0	9
132	0	10	2	1	5	0	0	0	18
136	4	12	3	0	0	0	0	0	19
140	8	7	0	0	0	1	0	0	16
Total	17	50	16	4	11	6	1	2	107

Captions to Figures

Figure 1: Smallest spherical, toroidal, Klein-bottle and elliptic fullerenes. The first column lists the graphs drawn in the plane, the second the map in the appropriate surface and the third the dual in the same surface. The examples are (a) the dodecahedron (dual = icosahedron), (b) the Heawood graph (dual = K_7), (c) a small Klein-bottle polyhex (dual = $K_{3,3,3}$) and (d) the Petersen graph (dual = K_6).

Figure 2: Examples of small polyhedral toroidal polyhexes presented as benzenoids with glued vertices (indicated on the periphery). These are the polyhexes with Coxeter/Goldberg codes $(2,1)$, $(3,0)$, $(2,2)$, $(3,1)$. The map $(2,1)$ is the Heawood graph, which is drawn in two different presentations in Fig. 1.

Figure 3: A tiling of the plane with combinatorial pentagons alone in which all vertices are of degree 3 or 4.

Figure 4: The smallest space fullerene: an assembly of 20- and 24-vertex spherical fullerenes in ratio 2 : 6 (see Wells [28]). Different metric variations of this structure appear as a clathrate, and as the best Kelvin foam. Its dual is A15, the structure of β-tungsten and Cr_3Si.

Figure 5: Leapfrogs of the the smallest polyhedral fullerenes on the surfaces S^2, T^2, K^2 and P^2.

Figure 6: Construction of a non-orientable surface such as a Möbius strip by twisting and gluing a planar system brings together p orbitals of opposite phase across the seam of the twist.

Figure 7: Construction of toroidal double covers of Klein-bottle polyhexes: (a) shows a bipartite K^2-polyhex on 18 vertices which is covered by a centrosymmetric D_{3d} torus on 36 vertices; (b) shows a non-bipartite K^2-polyhex on 24 vertices, covered by a 48-vertex, centrosymmetric D_{6h} torus. The steps in the construction are: (i) copy the Klein-bottle polyhex by a simple translation; (ii) flip the second copy by 180° about the translation vector; (iii) fuse the two copies. The arrows indicate edge identifications, and members of a pair of covering vertices are marked with the same symbol (filled circle or square). Note that in case (a), the two covering vertices are to be found in the same partite set on the torus, whereas in case (b) they are not.

Figure 8: Origin of zero eigenvalues in toroidal and Klein-bottle leapfrog polyhexes. An 18-vertex polyhex is shown in (a), numbered as for connection either as a T^2-fullerene or on K^2 with the twist occurring on gluing left and right edges of the parallelogram. This

polyhex is a leapfrog, as shown in (b) where shading identifies Clar faces and a double bond a Fries edge. Edges crossing the seam of the twist and therefore to be weighted by -1 in the Hückel problem on K^2 are in bold. When the polyhex is connected on T^2 all four vectors (c) to (f) represent eigenvectors of $\mathbf{A}$ with zero eigenvalue. When it is connected on K^2, (c) and (e) are zero-eigenvalue eigenvectors for $\mathbf{A}$ and (d) and (f) are zero-eigenvalue eigenvectors for $\mathbf{H}$, the weighted adjacency matrix. Details of the construction of (c) to (f) from local bonds and anti-bonds on Fries edges are given in ref. [43].

Figure 9: Multiple leapfrogs of a K^2-polyhex. The parent 18-vertex graph (thick lines) is leapfrogged to 54 and then to 162 vertices (thin lines). Only the first and third graphs are bipartite: when the patch is glued as indicated by the arrows black vertices give a consistent partite set (filled circles) in the parent and double leapfrog but not in the single leapfrog

FIGURE 1
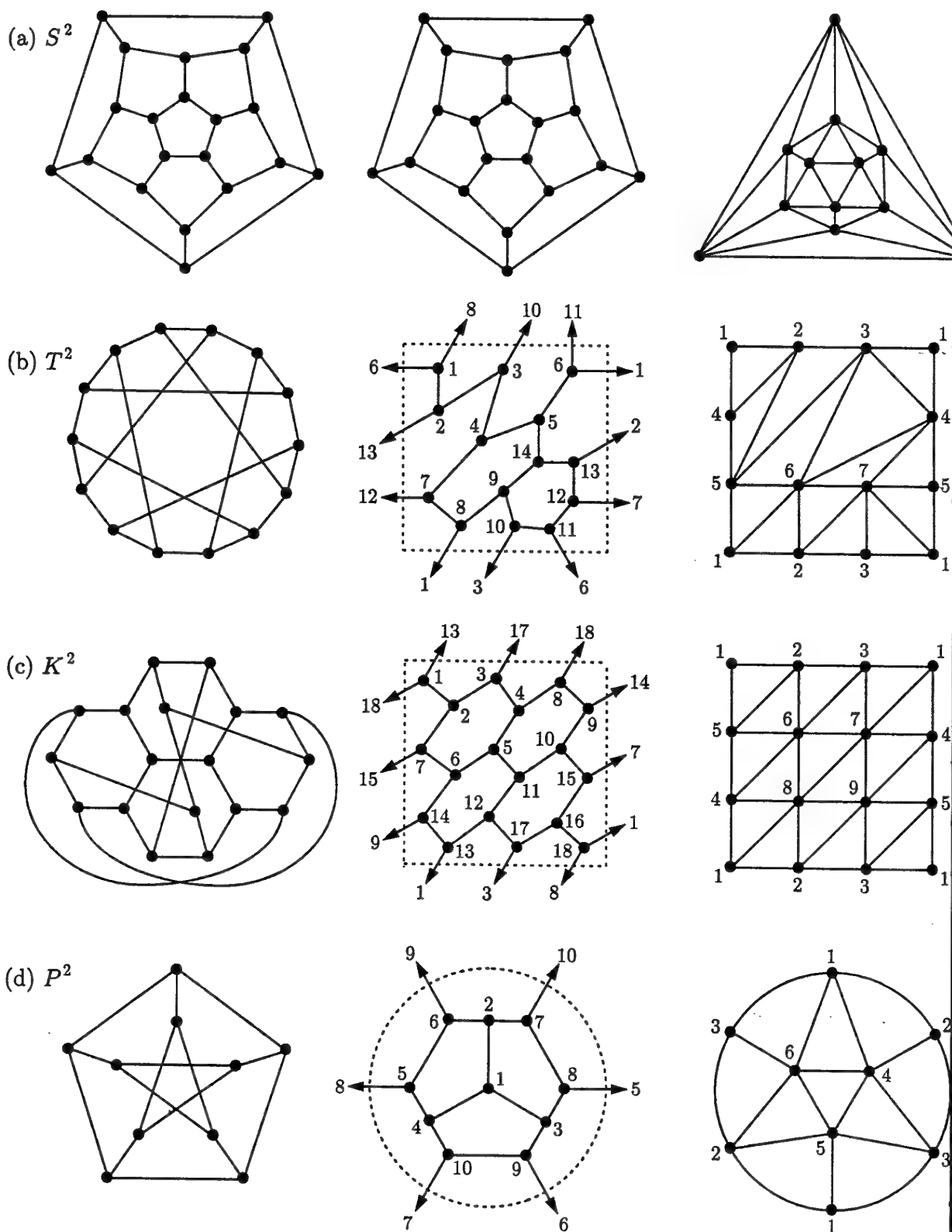
FIGURE 2
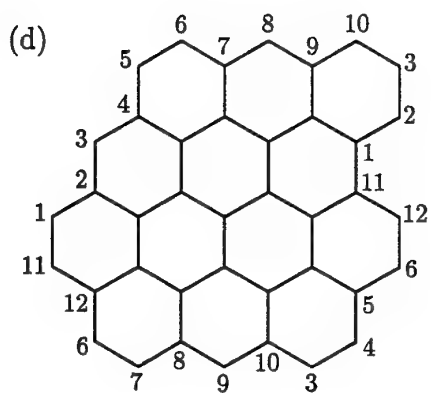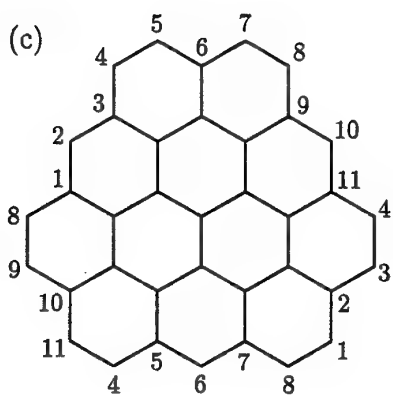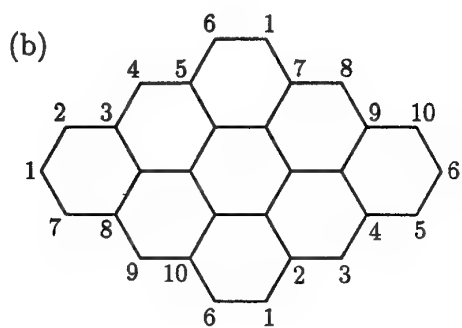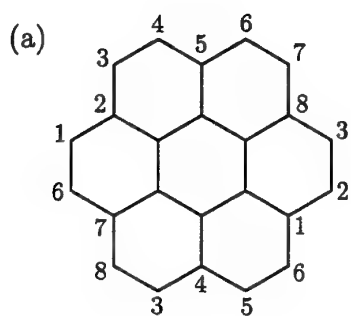
FIGURE 3
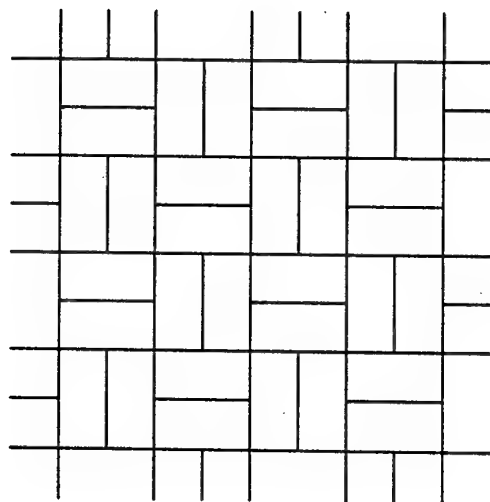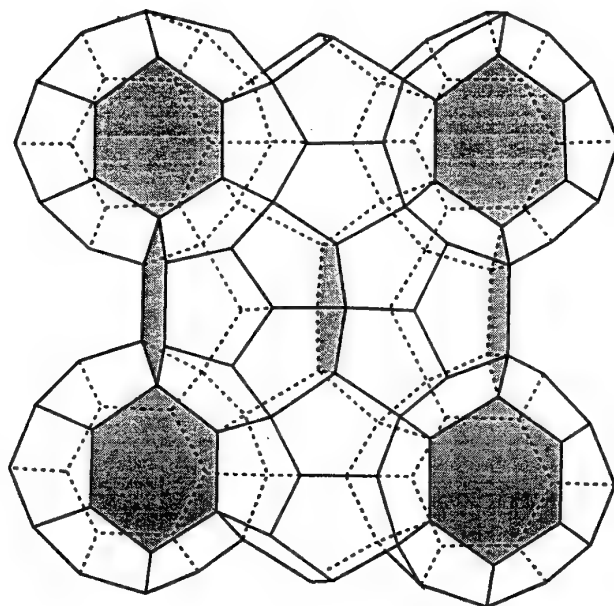
FIGURE 4

FIGURE 5

(a) S^2
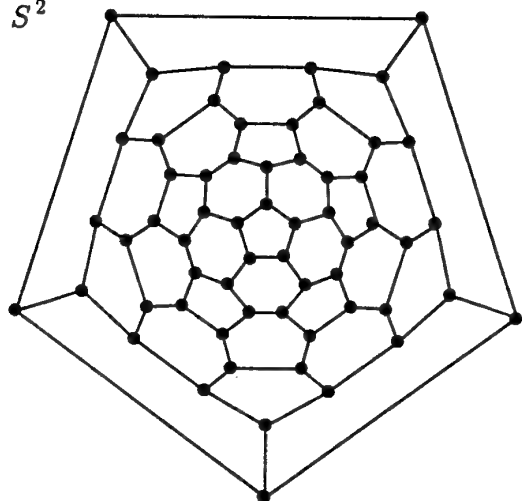
(b) T^2
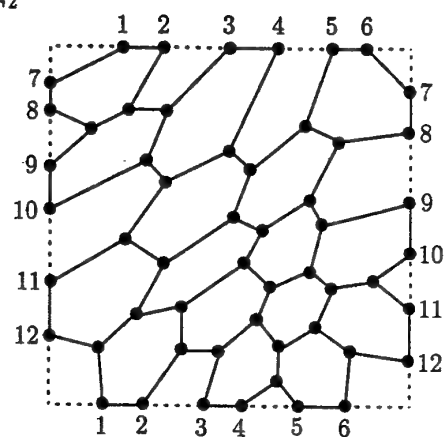
(c) K^2
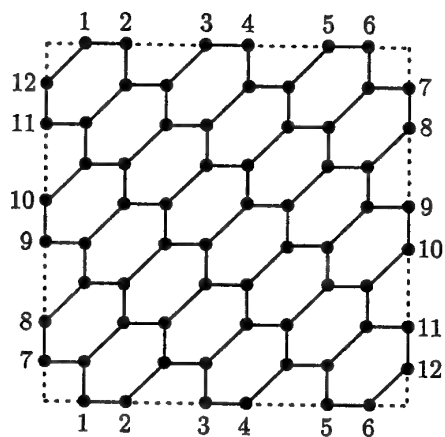
(d) P^2
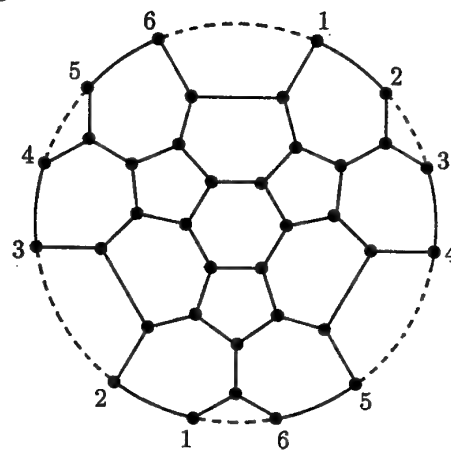
FIGURE 6
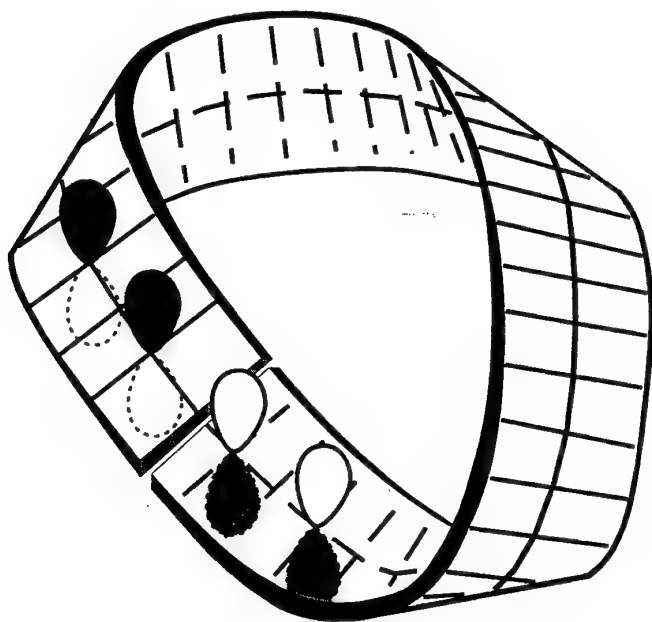
FIGURE 8
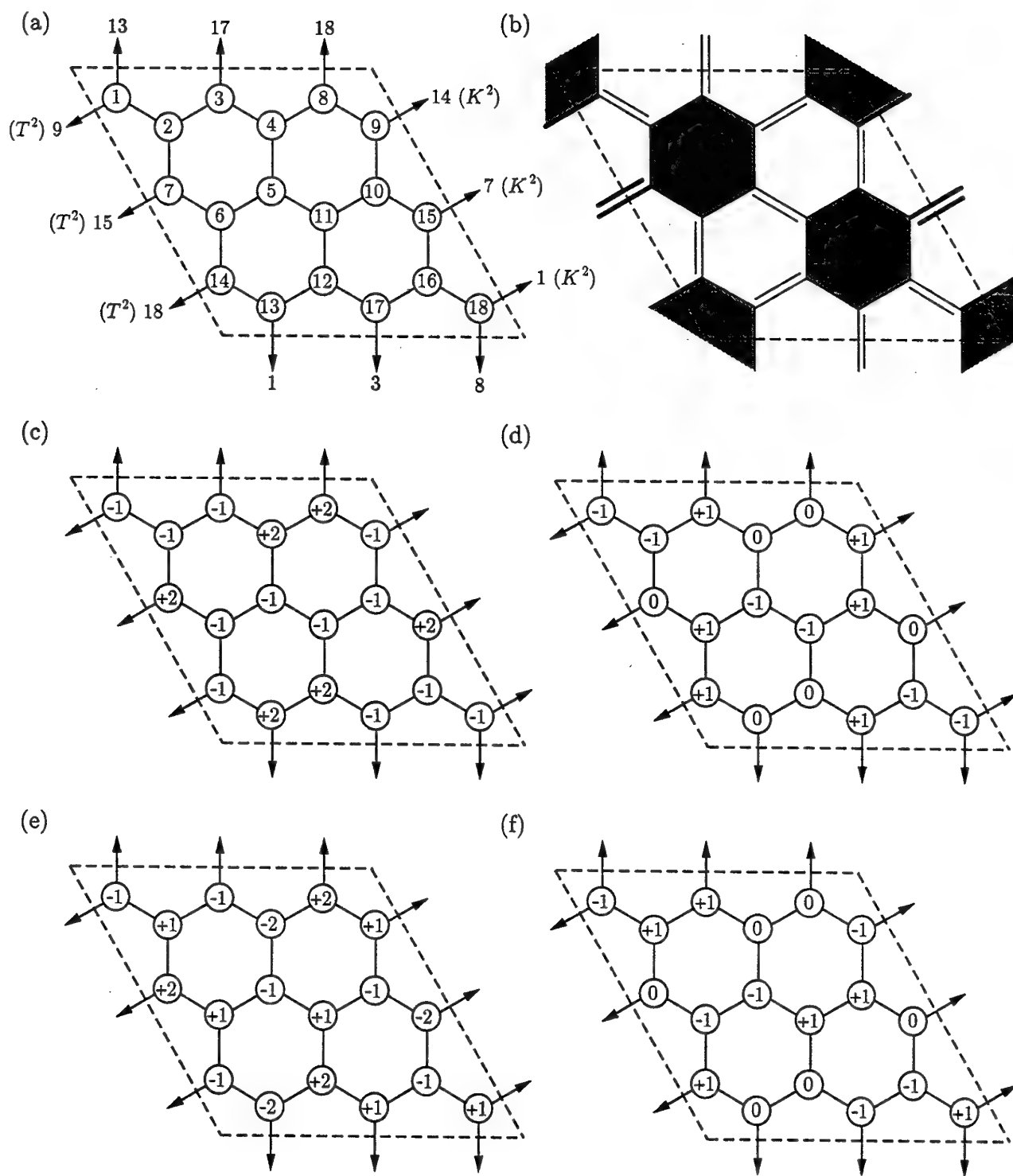
FIGURE 9
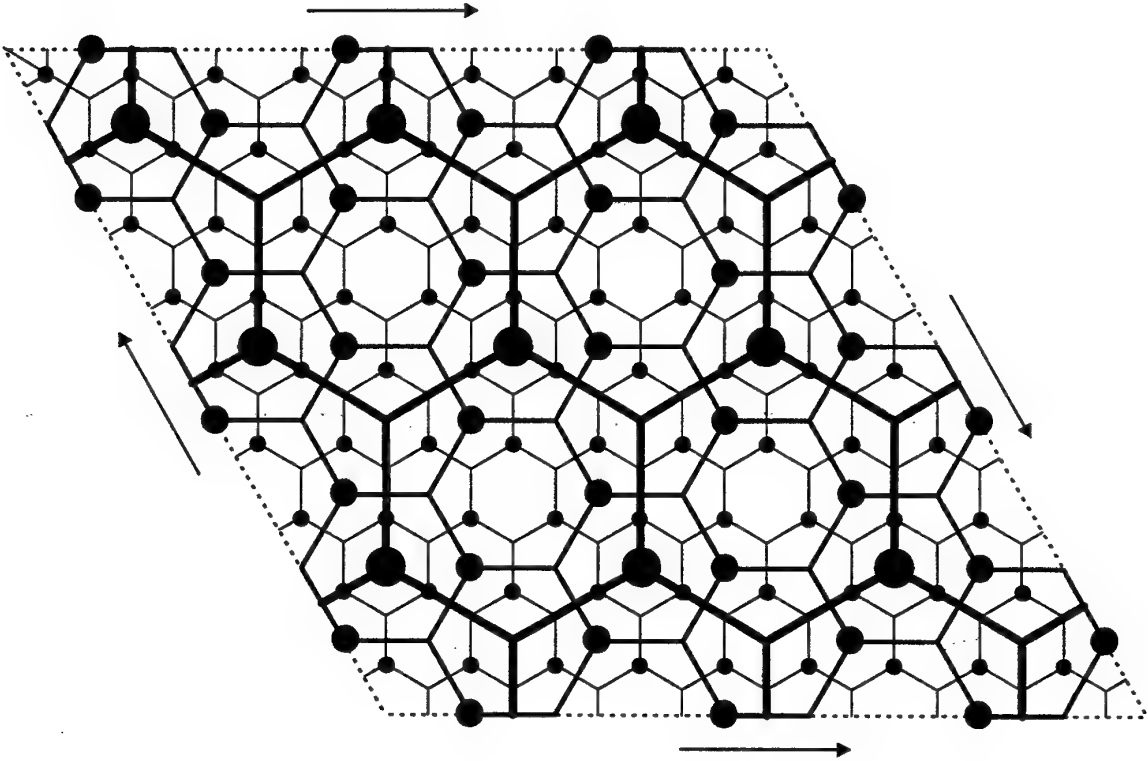
THE BYTE STRUCTURE VARIABLE LENGTH CODING (BS-VLC)

a new specific algorithm applied in the compression of trajectories generated by molecular dynamics

André Melo[1*], André T. Puga[2], Fernanda Gentil[1], Nelson Brito[1], Artur P. Alves[2] and Maria João Ramos[1]

[1]*CEQUP/Departamento de Química, Faculdade de Ciências do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal*

[2]*FEUP/INESC-Porto, Largo Mompilher 22, 4007 Porto Codex, Portugal.*

* *Author to whom all correspondence should be addressed.*

Phone number: 351-2-6082806

Fax number: 351-2-6082959

e-mail address: asmelo@fc.up.pt

1. INTRODUCTION

Molecular Dynamics (MD) has been applied to biomolecular systems and has shown to be able of elucidating on their energetic, dynamical and statistical mechanical properties. MD is a deterministic procedure in which the atoms in a molecule move according to classical mechanics. Thus, in a MD calculation we have to integrate Newton's equations of motion over time for the N atoms of the molecular system which is being studied

$$\partial^2 r(i,t)/\partial t^2 = m(i)^{-1} F(i,t) \quad i = 1,...,N \tag{1}$$

with $F(i,t) = -\partial V(r(1,t),...,r(N,t))/\partial r(i,t)$ as the force on atom i at time t, $V(r(1,t),...,r(N,t))$ as the potential energy function, $r(i,t)$ as the position of atom i at time t and $m(i)$ as the mass of atom i. In the present work, the leap-frog algorithm (van Gunsteren and Berendsen, 1990) has been used to compute the position vectors $r(i,t)$ using the forces and previous positions of the atoms at a series of time intervals which differ by Δt.

The set of atomic positions occupied in a given time t_j is called **conformation** (vector $r(i,t_j)$, i=1,...,N) and a succession of conformations, in n time intervals, is named a **trajectory** (matrix $[r(i,t_j)]$, i=1,...,N j=1,...,n). For simplicity, we shall use $r(i,j)$, i=1,...,N to represent a conformation and $[r(i,j)]$, i=1,...,N j=1,...,n to represent a trajectory. A trajectory is thus given by the following matrix

$$r(1,1) \quad r(2,1) \quad ...... \quad r(N,1) \Rightarrow \quad \text{1st conformation}$$

$$r(1,2) \quad r(2,2) \quad ...... \quad r(N,2) \Rightarrow \quad \text{2nd conformation} \qquad (2)$$

$$. \qquad . \qquad ...... \qquad .$$

$$r(1,n) \quad r(2,n) \quad ...... \quad r(N,n) \Rightarrow \quad \text{nth conformation}$$

The trajectories generated by a MD calculation are the basis for the calculation of the system properties. However, the correspondent binary trajectory files are extensive and create problems of storage space. The classical lossless compression algorithms, such as the Huffman coding (Huffman, 1952) used in the compression pack utility, adaptive Huffman (Gallager, 1978) used in compact, LZW (Welch, 1984) used in compress and LZ77 (Ziv and Lempel, 1977) used in gzip, give poor efficiencies in the compression of this type of files. Therefore, specific lossy algorithms, which increase significantly the compression efficiency preserving a high degree of precision, are of great importance to attain a better approach to this problem.

This work introduces the reader to a new specific algorithm, named Byte Structure Variable Length Coding (BS-VLC), which increases significantly the compression efficiencies of the best classical lossless algorithms preserving a high degree of precision. This algorithm was used in the compression of trajectory files generated by MD applied to the biological systems trypsin and trypsin:PTI complex.

4

2. METHODS

2.1. COMPRESSION WITH THE BS-VLC ALGORITHM

The initial data in this process is the trajectory matrix $[r(i,j)]$, $i=1,...,N$ $j=1,...,n$ which in the cartesian space can be substituted by three matrices $[r_u(i,j)]$ with $u=x$, y or z respectively and $i=1,...,N$ $j=1,...,n$. The trajectories are usually stored in binary files with every coordinate figuring as a 4 bytes real number. All of this data is submitted to three steps in a BS-VLC compression: pre-processing, quantization and variable length coding. These will be analyzed separately.

Pre-processing   The proposed methodology here is based on the conversion of the initial data into differential trajectory matrices $[\Delta r_u(i,j)]$, $i=1,...,N$ $j=1,...,n$ in which the different components are given by coordinate differences, i.e.

$$\Delta r_u(i,j) = r_u(i,j) - r_u^{ref}(i,j) \qquad\qquad (3)$$

where $r_u^{ref}(i,j)$ is the reference coordinate associated with $r_u(i,j)$. It is necessary to store n_0 conformations (standard integral conformations ($[r_u^{\ominus}(i,k)]$, $i=1,...,N$ $j=1,...,n_0$) in its original form to allow the rebuilding of trajectory matrices $[r_u(i,j)]$ from the correspondent differential matrices $[\Delta r_u(i,j)]$, in the decompression. The number of standard integral conformations depends on the criteria used in the selection of the reference coordinates $r_u^{ref}(i,j)$ and will be discussed later in section 2.4.

Quantization This information is further transformed into integer differential trajectory matrices [u(i,j)], i=1,...,N j=1,...,n with components given by coordinate differences converted to integers multiplied by a scaling factor (Scale) as follows

$$u(i,j) = \Delta r_u(i,j) \times Scale \qquad\qquad (4)$$

The higher the scaling factor, the higher the precision of the process is and the lower the compression efficiency becomes; the contrary also applies, i.e., to increase the compression efficiency one can lower the scaling factor with the added disadvantage of lowering the precision. It is important to find a middle point which will provide a good compression coupled with a reasonable error.

Variable Length Coding (VLC) Subsequently, the elements of the integer differential matrices [u(i,j)] are compressed using a variable length code, where the trajectories are subdivided into sets (structures) which are separately coded.

2.2. CODING WITHIN THE BS-VLC ALGORITHM

Byte integer signal coding The integers, components of the integer differential trajectory matrices [u(i,j)], can be represented by implicit signal coding or explicit signal coding. The former is the conventional form of partitioning integers into bytes, i.e., 1 bit is always kept to represent the signal and the other N-1 bits represent the respective absolute value; in the latter, signal bits are explicitly codified and grouped into signal bytes. This representation can be less expensive if all the data is centered in

6

certain regions. In a 32 bit machine the minimum number of bytes necessary for the representation of absolute values, according to the interval where they are, is

$$1 \text{ byte} \iff [0 : 2^8 -1]$$
$$2 \text{ byte} \iff [2^8 : 2^{16} -1] \tag{5}$$
$$3 \text{ byte} \iff [2^{16} : 2^{24} -1]$$
$$4 \text{ byte} \iff [2^{24} : 2^{32} -1]$$

However, when we are working with signed integers, one half of this space has to be used to represent negative integers. Therefore, we end up with the following association between number of bytes and intervals

$$1 \text{ byte} \iff [0 : 2^7-1] \cup [-1 : -2^7]$$
$$2 \text{ byte} \iff [2^7 : 2^{15} -1] \cup [-2^7 -1: -2^{15}] \tag{6}$$
$$3 \text{ byte} \iff [2^{15} : 2^{23} -1] \cup [-2^{15} -1: -2^{23}]$$
$$4 \text{ byte} \iff [2^{23} : 2^{31} -1] \cup [-2^{23} -1: -2^{31}]$$

On the other hand, the signal can be explicitly coded allowing the absolute values to be represented according to eq. (5). One possible way to perform this is to subdivide the data in groups of eight elements and to associate the value 1 to their signals when they are positive and the value 0 when they are negative.

For example: + - + + - - + + $\iff$ 1 0 1 1 0 0 1 1

Each group of eight signal bits is then stored in a signal byte and the entire sequence of signal bytes is stored in a signal vector. The dimension (nsbyte) of this vector is equal to the number of coordinates over 8. Associating eqs. (5) and (6), the 32 bit integer space can be subdivided into seven intervals which need different minimum numbers of bytes for coding according to whether implicit or explicit signal coding is used. All this information has been collected in Table I. It can be observed from the table that explicit signal coding is favored when data is concentrated in intervals 2, 4 or 6.

Concept of structure   Structures are sets within the integer differential trajectory matrices, where elements exhibit some sort of correlation. BS-VLC is based on the fact that a structure could be represented by fewer bytes than entire trajectory matrices. All the elements within a structure (S_i) are represented by the number of bytes ($nbmax(S_i)$) necessary to codify the element with largest absolute value. However, it is necessary to keep one byte extra for each structure to indicate the number of bytes used in its representation. If a trajectory matrix is subdivided into n_0 standard integral conformations, n_1 structures that use implicit signal coding and n_2 structures that use explicit signal coding, the byte length of a compressed file (CFBL) is given by

$$CFBL = ovh + (n_1 + n_2) + nsbyte + \sum_{i=1}^{n1} n(S_i) \times nbmax1(S_i) + \sum_{j=1}^{n2} n(S_j) \times nbmax2(S_j) + ns_0 \quad (7)$$

In eq. (7), ovh is the structure organization overhead which indicates how the trajectory file was compressed. This information is necessary to the decompression process, but represents a very insignificant part (less than 0.01%) of CFBL. In the same

8

equation, $n(S_i)$ and $n(S_j)$ are the number of elements of structures S_i and S_j, respectively; $nbmax1(S_i)$ is the number of bytes necessary to represent the element of structure S_i with largest absolute value determined using the third column of Table I; $nbmax2(S_j)$ is the number of bytes necessary to represent the element of structure S_j with largest absolute value determined using the fourth column of Table I, ns_0 is the number of bytes necessary to store the standard integral conformation and nsbyte is the total number of elements represented with explicit signal coding over 8:

$$ns_0 = 4 \times 3N \times n_0 \tag{8}$$

$$nsbyte = \sum_{j=1}^{n2} n(S_j)/8 \tag{9}$$

Types of structures The BS-VLC algorithm considers two types of structures: structures associated with temporal correlations (blocks and conformations within a block) and structures associated with spatial correlations (atoms within a block and atomic cartesian coordinates within a block).

Block The block structure is the largest one considered in the BS-VLC method. A block is a set of sequential conformations and this is a natural structure with the pre-processing and quantization methodologies adopted in this work. If implicit signal coding is assumed, the byte length of block B_i ($BBL(B_i)$) is given by

$$BBL(B_i) = 3N \times nconf(B_i) \times nbmax1(B_i) \tag{10}$$

9

where $\mathrm{nconf}(B_i)$ is the number of conformations of block B_i. If explicit signal coding is assumed, $\mathrm{BBL}(B_i)$ is given by

$$\mathrm{BBL}(B_i)= \mathrm{nsbyte}(B_i) + 3N \times \mathrm{nconf}(B_i) \times \mathrm{nbmax2}(B_i) \qquad (11)$$

where $\mathrm{nsbyte}\,(B_i)$ is calculated as

$$\mathrm{nsbyte}(B_i)=\mathrm{nconf}(B_i) \times 3N \,/\, 8 \qquad (12)$$

Conformation within a block The conformation within a block is the smallest temporal structure considered in the BS-VLC method. In this case, $\mathrm{BBL}(B_i)$ with implicit signal coding is given by

$$\mathrm{BBL}(B_i)= \mathrm{nconf}(B_i) + 3N \sum_{k=1}^{nconf(Bi)} \mathrm{nbmax1}(\mathrm{conf}(k)) \qquad (13)$$

If explicit signal coding is assumed, $\mathrm{BBL}(B_i)$ is calculated as

$$\mathrm{BBL}(B_i)= \mathrm{nconf}(B_i) + \mathrm{nsbyte}(B_i) + 3N \sum_{k=1}^{nconf(Bi)} \mathrm{nbmax2}(\mathrm{conf}(k)) \qquad (14)$$

Atom within a block   In an MD trajectory it is possible to establish spatial correlation. An atom within a block is a structure that reflects this type of correlation. For example, in a protein the differential coordinates of the side chain atoms are usually larger than the differential coordinates of the main chain atoms. In this case, $\mathrm{BBL}(B_i)$ with implicit signal coding is given by

10

$$BBL(B_i) = N + 3 \, nconf(B_i) \sum_{k=1}^{N} nbmax1(at(k)) \tag{15}$$

If explicit signal coding is assumed, $BBL(B_i)$ is calculated as

$$BBL(B_i) = N + nsbyte(B_i) + 3nconf(B_i) \sum_{k=1}^{N} nbmax2(at(k)) \tag{16}$$

Atomic cartesian coordinate within a block   The atomic cartesian coordinate within a block is a subdivision of the atom structure and it is appropriate when the potential energy function exhibits some type of anisotropy. Here, $BBL(B_i)$ with implicit signal coding, is given by

$$BBL(B_i) = 3N + nconf(B_i)\{ \sum_{k=1}^{N} nbmax1(x(k)) +$$

$$+ \sum_{k=1}^{N} nbmax1(y(k)) + \sum_{k=1}^{N} nbmax1(z(k))\} \tag{17}$$

If explicit signal coding is assumed, $BBL(B_i)$ is calculated as

$$BBL(B_i) = 3N + nsbyte(B_i) +$$

$$+ nconf(B_i) \{ \sum_{k=1}^{N} nbmax2(x(k)) + \sum_{k=1}^{N} nbmax2(y(k)) + \sum_{k=1}^{N} nbmax2(z(k))\} \tag{18}$$

11

In eqs. (17) and (18) x(k), y(k) and z(k) are the cartesian coordinates x,y and z of atom k in block B_i, respectively.

2.3. DECOMPRESSING THE FILES OBTAINED WITH THE BS-VLC ALGORITHM

A BS-VLC decompression of files follows three steps which will be analyzed separately, namely variable length decoding, inverse quantization and post-processing.

Variable Length Decoding (VLD) In this step, the compressed file is decoded restoring the original integer differential matrices $[\mathbf{u}(i,j)]$.

Inverse quantization The differential matrices $[\mathbf{u}(i,j)]$ are used to produce the decompressed differential matrices $[\Delta \tilde{\mathbf{r}}_u (i,j)]$ with

$$\Delta \tilde{\mathbf{r}}_u (i,j) = \mathbf{u}(i,j)/Scale \qquad (19)$$

The elements of the decompressed differential matrices differ from the original differential matrices from a given quantity $e(\Delta \mathbf{r}_u(i,j))$:

$$\Delta \tilde{\mathbf{r}}_u (i,j) = \Delta \mathbf{r}_u(i,j) + e(\Delta \mathbf{r}_u(i,j)) \qquad (20)$$

The magnitude of the error $e(\Delta \mathbf{r}_u(i,j))$ depends on the scaling factor selected.

12

Post-processing   The decompressed trajectory matrices $[\,\tilde{\mathbf{r}}_u(i,j)]$ are then calculated from the decompressed differential trajectory matrices $[\Delta\,\tilde{\mathbf{r}}_u(i,j)]$ and from the decompressed reference coordinates $\tilde{\mathbf{r}}_u^{ref}(i,j)$.

$$\tilde{\mathbf{r}}_u(i,j) = \tilde{\mathbf{r}}_u^{ref}(i,j) + \Delta\,\tilde{\mathbf{r}}_u(i,j) \tag{21}$$

The elements of the decompressed trajectory matrices differ from the elements of the original trajectory matrices by a quantity $e(\mathbf{r}_u(i,j))$.

$$\tilde{\mathbf{r}}_u(i,j) = \mathbf{r}_u(i,j) + e(\mathbf{r}_u(i,j)) \tag{22}$$

The characteristics of the errors $e(\mathbf{r}_u(i,j))$ depend on the type of the coordinates $\mathbf{r}_u^{ref}(i,j)$ selected. The nature of these dependencies will be discussed later in section 2.4. A flow chart of the BS-VLC algorithm is given in Figure 1.

2.4. SELECTION OF THE REFERENCE COORDINATES

Several alternative selection criteria are possible to choose the reference coordinates $\mathbf{r}_u^{ref}(i,j)$ used in eq. (3). The characteristics of the errors $e(\mathbf{r}_u(i,j)$, in eq. (22), are conditioned by the reference coordinates selected. We have considered the following types of reference coordinates:

Atomic coordinates The reference coordinates are defined as the atomic coordinates immediately before the present $\mathbf{r_u}(i,j)$ coordinates

$$\mathbf{r_u^{ref}}(i,j) = \mathbf{r_u}(i,j-1) \tag{23}$$

The standard integral conformation is the first conformation of the trajectories $\mathbf{r_u}(i,1)$, with the necessary number of bytes to be stored, given as

$$ns_0 = 4 \times 3 \text{ na} \tag{24}$$

Equation (21) becomes

$$\tilde{\mathbf{r}}_u(i,j) = \tilde{\mathbf{r}}_u(i,j-1) + \Delta \tilde{\mathbf{r}}_u(i,j) \tag{25}$$

Substituting eq. (20) into eq. (25), one obtains

$$\tilde{\mathbf{r}}_u(i,j) = \tilde{\mathbf{r}}_u(i,j-1) + \Delta \mathbf{r_u}(i,j) + e(\Delta \mathbf{r_u}(i,j)) \tag{26}$$

Repeating this substitution for all the $\tilde{\mathbf{r}}_u(i,j-1)$ we obtain

$$\tilde{\mathbf{r}}_u(i,j) = \mathbf{r_u}(i,1) + \sum_{k=2}^{j} \Delta \mathbf{r_u}(i,k) + \sum_{k=2}^{j} e(\Delta \mathbf{r_u}(i,k)) \tag{27},$$

or

$$\tilde{r}_u(i,j) = r_u(i,j) + \sum_{k=2}^{j} e(\Delta r_u(i,k)) \tag{28}$$

Comparing eqs. (28) and (22), the errors $e(r_u(i,j)$ can be calculated as

$$e(r_u(i,j)) = \sum_{k=2}^{j} e(\Delta r_u(i,k)) \tag{29}$$

If the successive values $e(\Delta r_u(i,j))$ are not well compensated this will originate cumulative errors $e(r_u(i,j))$.

Atomic coordinates within a block   One possible methodology which partially prevents the cumulative nature of the error inherent to the previous selection, consists in truncating its accumulation at the end of each block. In this situation, it is necessary to store, not only the first conformation of the trajectory, but also the first conformation of all the blocks. Consequently, the error, eqs. (26)-(28), is still cumulative within a block but becomes null for the transition between the last conformation of a block and the first conformation of the next block. The compression efficiencies are also lightly reduced by this procedure:

$$ns_0 = 4 \times 3N \times nblo \tag{30}$$

where nblo is the total number of blocks used.

Decompressed atomic coordinates   Here, the reference coordinates are defined as

15

$$r_u^{ref}(i,j) = \tilde{r}_u(i,j-1) \qquad (31)$$

Equation (3) becomes,

$$\Delta r_u(i,j) = r_u(i,j) - \tilde{r}_u(i,j-1) \qquad (32)$$

and eq. (20) can be rewritten as

$$\Delta \tilde{r}_u(i,j) = r_u(i,j) - \tilde{r}_u(i,j-1) + e(\Delta r_u(i,j)) \qquad (33)$$

Equation (33) can be rearranged as

$$\tilde{r}_u(i,j-1) + \Delta \tilde{r}_u(i,j) = r_u(i,j) + e(\Delta r_u(i,j)) \qquad (34)$$

or

$$\tilde{r}_u(i,j) = r_u(i,j) + e(\Delta r_u(i,j)) \qquad (35)$$

Comparing eqs. (35) and (22), we conclude that the errors $e(r_u(i,j))$ can be calculated as

$$e(r_u(i,j)) = e(\Delta r_u(i,j)) \qquad (36)$$

Here, the error becomes non cumulative preserving the compression efficiency obtained with reference coordinates a). This formulation corresponds to the classical

16

scheme DPCM (Differential Pulse Code Modulation) (Jayan and Noll, 1984) which is the basis of several audio, image and video compression algorithms such as JPEG (Wallace, 1991) and MPEG (ISO/IEC, 1994).

The best way to evaluate the cumulative characteristics of the error associated with the compression/decompression process is to represent the conformational root mean square deviation (rms) between the decompressed $\tilde{r}_u(i,j)$ and the original coordinates $r_u(i,j)$,

$$rms = \{[\sum_{i=1}^{N} (\tilde{r}_x(i,j) - r_x(i,j))^2 + (\tilde{r}_y(i,j) - r_y(i,j))^2 + (\tilde{r}_z(i,j) - r_z(i,j))^2]/(3N-1)\}^{1/2} \quad (37),$$

as a function of simulation time.

The total mean square deviation ($\overline{rms}$) is given by:

$$\overline{rms} = \{[\sum_{j=1}^{n}\sum_{i=1}^{N} (\tilde{r}_x(i,j)-r_x(i,j))^2 + (\tilde{r}_y(i,j)-r_y(i,j))^2 + (\tilde{r}_z(i,j)-r_z(i,j))^2]/(3N \times n-1)\}^{1/2} \quad (38)$$

allowing an evaluation of the global precision of the compression/decompression process.

2.5. IMPLEMENTATION OF THE BS-VLC ALGORITHM

In the BS-VLC compression scheme the following procedure has to be considered

Pre-processing

1) Read initial data $[r_u(i,j)]$.

2) Select the number of blocks (nblo).

3) Select the reference coordinates (atomic coordinates, atomic coordinate within a block or decompressed atomic coordinates).

4) Store the standard integral conformations $[r_u^\oplus(i,k)]$.

5) Calculate the differential trajectory matrices $[\Delta r_u(i,j)]$.

Quantization

1) Select the scaling factor (Scale).

2) Calculate the integer differential matrices $[u(i,j)]$.

VLC

For each block:

1) Select the structure (entire block, conformation within the block, atom within the block or atomic cartesian coordinate within the block) and the signal coding (explicit or implicit) which will allow a more efficient compression of the block (eqs. (10)-(18)).

2) Compress the block using the structure and the signal coding selected in 1).

The byte length of the compressed file (CFBL) is given by

$$CFBL = ovh + \sum_{j=1}^{nblo} BBL(B_i) + ns_0 \qquad (39)$$

In the BS-VLC decompression scheme the following procedure must to be considered:

VLD

For each block:

1) Read the structure (entire block, conformation within the block, atom within the block or atomic cartesian coordinate within the block) and the signal coding (explicit or implicit) used in the compression.

2) Decompress the block using the structure and signal coding read in 1) (the integer differential matrices $[u(i,j)]$ are rebuilt).

Inverse quantization

1) Read the scaling factor (Scale).

2) Calculate the decompressed differential matrices $[\Delta \tilde{r}_u (i,j)]$.

Post-processing

1) Read the standard integral conformations $[r_u^{\ominus}(i,k)]$.

2) Calculate the decompressed trajectory matrices $[\tilde{r}_u (i,j)]$.

3. RESULTS AND CONCLUSIONS

In this work two solvated proteic systems, trypsin and trypsin:PTI complex, have been studied. Trypsin is a digestive enzyme and PTI (Pancreatic Trypsin Inhibitor) is a natural inhibitor of trypsin.

In the MD simulations (Melo and Ramos, 1997), all the water molecules and amino acid residues within a 15Å sphere, centered in the active center of trypsin, were allowed to move. Harmonic forces were used to restrain any water molecule from leaving the 15-18 Å boundary; this was achieved by constraining the oxygen atoms of the water molecules to their initial positions using a force constant of 0.6 kcalmol^{-1}Å^{-1}. The other residues were included in the determination of the energy and forces, but were kept fixed in their starting positions. The choice for performing the MD calculation in this way had its reasons in the size of the simulation which would have been prohibitive otherwise (Melo and Ramos, 1997). A total number of 1815 atoms for solvated trypsin and 1666 for solvated trypsin:PTI were allowed to move during the MD simulations. Newton's equations of motion were integrated every 0.001ps using the leap-frog algorithm (van Gunsteren and Berendsen, 1990) and 120 ps trajectories were generated for both systems studied. Total simulation time was 6 240 ns for trypsin, 6 400 ns for trypsin:PTI and 12 640 ns for both trypsin and trypsin:PTI. The trajectory matrices have the following dimensions:

$\mathbf{r_u}$(1815,6000) $\Rightarrow$ for solvated trypsin

$\mathbf{r_u}$(1666,6000) $\Rightarrow$ for solvated trypsin:PTI

All simulations were carried out with the program CHARMM (Brooks *et al.*, 1983) and the two trajectories were compressed using the lossless compressors (compress, gzip, pack and compact) as well as the BS-VLC algorithm.

In the BS-VLC compressions, a total number of 60 blocks was used. Preliminary compressions, using the three alternative reference coordinates (atomic coordinates, atomic coordinates within a block and decompressed atomic coordinates) and a scaling factor of 10^5, were performed. The conformational root mean square deviation (rms) was computed as a function of simulation time. The results obtained are presented in Figure 2; they confirm that atomic coordinates, atomic coordinates within a block and decompressed atomic coordinates lead to a cumulative, a truncated cumulative and a non cumulative error, respectively. Consequently, as has been pointed in section 5, the decompressed atomic coordinates are the most appropriate selection for reference coordinates. This selection was used in the remainder BS-VLC compression presented here.

To evaluate the efficiency of the BS-VLC algorithm, several compressions of trypsin and trypsin:PTI trajectory files were performed, using 28 different values as scaling factors. The total root mean square deviations ($\overline{\text{rms}}$) were computed for all the cases. The results obtained are presented in Table II and Figure 3. Additionally, the compression efficiencies, obtained with different algorithms, can be visualized in Figure 4.

The analysis of both Table II, Figure 3 and Figure 4 is extremely favorable to the BS-VLC algorithm. In fact, when a scaling factor of 10^7 is used BS-VLC algorithm has a lossless behavior and presents a significant larger compression efficiency than the classical lossles algorithms.

21

Green et al. (1995) have made available a compression algorithm which reaches 70% as compared to 75% achieved in this work. Additionally, this algorithm leads to a cumulative error in the compression/decompression process while BS-VLC has non cumulative error behavior.

Here, all the obtained results enable us to conclude that BS-VLC has near lossless behavior ($\overline{rms}=0$) when a scaling factor close to 10^7 is used. In this situation, BS-VLC nearly triplicates the compression efficiency of the best classical lossless algorithm (LZ77 used in gzip). In addition, larger compression efficiencies ($\approx50\%$) can be managed with BS-VLC preserving a high degree of precision ($\overline{rms}$ between 10^{-5} and 10^{-6}). For compression efficiencies larger than 50%, the precision decreases significantly. However, a compression with the maximum efficiency possible (75%) within this algorithm can be performed with good precision ($\overline{rms}\approx10^{-3}$).

REFERENCES

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) *J. Comput. Chem.*, **4**, 187.

Gallager, R.G. (1978) *IEEE Trans. Inform. Theory*, **IT-24**, 668.

Green D. G., Meacham, K.E., Surridge, M., van Hoesel, F. and Berendsen H.J.C. (1995) *Metecc-95 Proceedings*, 434.

Huffman, D.V. (1952) *Proc. I.R.E.*, **40**, 1098.

ISO/IEC JTC 1/SC 29/WG 11, *ISO/IEC CD 13818-3: Information Technology - Generic Coding of Moving Pictures and Associated Audio Information*, 1994.

Jayan, N.S. and Noll, P. (1984) *Digital Coding of Waveforms; Principles and Applications to Speech and Video*, Prentice-Hall, Signal Processing Series, Englewood Cliffs.

Melo, A. and Ramos, M.J. (1997) *J. Peptide Res.*, **50**, 382.

Wallace, G.K. (1991) *Comm. ACM*, **34**, 30.

Welch, T.A. (1984) *IEEE Computer*, **17**, 8.

van Gunsteren, W.F. and Berendsen, H.J.C. (1990) *Angew. Chem., Int. Ed. Eng.*, **29**, 98.

Ziv, J. and Lempel, A. (1977) *IEEE Trans. Inform. Theory*, **IT-23**, 337.

FIGURE CAPTIONS

FIGURE 1. Flow chart of the BS-VLC algorithm. The following steps have to be considered:

Compression- (1) Pre-processing; (2) Quantization; (3) VLC. Decompression- (1') VLD; (2') Inverse quantization; (3') Post-processing.

FIGURE 2. Conformational root mean square deviations (rms) between the original trajectories, obtained by molecular dynamics applied to (A) trypsin and (B) trypsin:PTI complex, and the trajectories obtained by sequential BS-VLC compression and decompression with a scaling factor of 10^5 as function of the simulation time (t). Three alternative reference coordinates, atomic coordinates (---), atomic coordinates within a block (...) and decompressed atomic coordinates (——), were used.

FIGURE 3. Total root mean square deviations ($\overline{rms}$) between the original trajectories, obtained by molecular dynamics applied to (A) trypsin and (B) trypsin:PTI complex, and the trajectories obtained by sequential BS-VLC compression and decompression with 28 different scaling factors as function of compression efficiency.

FIGURE 4. Byte length (BL) of initial trajectory files, obtained by molecular dynamics applied to trypsin and trypsin:PTI complex, and of compressed files obtained using different algorithms. The compression efficiencies and total root mean square deviations, between the original trajectory and the trajectories obtained by sequential compression and decompression, are also indicated within round and square brackets, respectively.

Compressed file

3

1'

Standard integral conformations

$r_x^\theta(i,k)$ $r_y^\theta(i,k)$ $r_z^\theta(i,k)$

Integer differential trajectory matrices

$x(i,j)$ $y(i,j)$ $z(i,j)$

2

2'

Standard integral conformations

$r_x^\theta(i,k)$ $r_y^\theta(i,k)$ $r_z^\theta(i,k)$

Differential trajectory matrices

$\Delta r_x(i,j)$ $\Delta r_y(i,j)$ $\Delta r_z(i,j)$

Standard integral conformations

$r_x^\theta(i,k)$ $r_y^\theta(i,k)$ $r_z^\theta(i,k)$

Decompressed differential trajectory matrices

$\Delta \widetilde{r}_x(i,j)$ $\Delta \widetilde{r}_y(i,j)$ $\Delta \widetilde{r}_z(i,j)$

1

Trajectory matrices

$r_x(i,j)$ $r_y(i,j)$ $r_z(i,j)$

3'

Decompressed trajectory matrices

$\widetilde{r}_x(i,j)$ $\widetilde{r}_y(i,j)$ $\widetilde{r}_z(i,j)$

FIGURE 1.

FIGURE 2.

FIGURE 3.

FIGURE 4.

Table I: Minimum number of byte coding for different intervals of integers numbers.

interval identification	interval	Minimum no. of byte coding	
		implicit signal coding	explicit signal coding
1	$[0:2^7-1]\cup[-1:-2^7]$	1	1
2	$[2^7:2^8-1]\cup[-2^7-1:-2^8+1]$	2	1
3	$[2^8:2^{15}-1]\cup[-2^8:-2^{15}]$	2	2
4	$[2^{15}:2^{16}-1]\cup[-2^{15}-1:-2^{16}+1]$	3	2
5	$[2^{16}:2^{23}-1]\cup[-2^{16}:-2^{23}]$	3	3
6	$[2^{23}:2^{24}-1]\cup[-2^{23}-1:-2^{24}+1]$	4	3
7	$[2^{24}:2^{32}-1]\cup[-2^{24}:-2^{31}]$	4	4

Table II. Compression efficiencies and total root mean square deviations ($\overline{rms}$) between the original trajectories, obtained by molecular dynamics applied to trypsin and trypsin:BPTI complex, and the trajectories obtained by sequential BS-VLC compression and decompression with a scaling factor (Scale).

Trypsin			Trypsin:BPTI		
Scale	Compression efficiency (%)	$\overline{rms}$ (Å)	Scale	Compression efficiency (%)	$\overline{rms}$ (Å)
1.000×10^7	24.7	0.000	1.000×10^7	24.7	0.000
9.400×10^6	24.8	6.000×10^{-9}	9.400×10^6	24.9	6.000×10^{-9}
8.850×10^6	24.9	7.000×10^{-9}	8.300×10^6	25.0	9.000×10^{-9}
8.300×10^6	25.0	7.100×10^{-9}	7.200×10^5	25.1	1.400×10^{-8}
7.200×10^5	25.1	1.000×10^{-8}	4.108×10^5	26.3	3.664×10^{-7}
3.664×10^5	27.3	8.140×10^{-7}	3.664×10^5	29.2	6.090×10^{-7}
3.220×10^5	30.2	9.290×10^{-7}	3.220×10^5	32.3	7.790×10^{-7}
2.776×10^5	32.3	1.055×10^{-6}	2.776×10^5	34.6	9.050×10^{-7}
2.332×10^5	33.5	1.321×10^{-6}	2.332×10^5	35.6	1.055×10^{-6}
1.888×10^5	34.7	1.653×10^{-6}	1.888×10^5	36.7	1.285×10^{-6}
1.444×10^5	38.1	2.182×10^{-6}	1.444×10^5	39.4	1.685×10^{-6}
1.000×10^5	45.0	2.923×10^{-6}	1.000×10^5	45.2	3.197×10^{-6}
9.400×10^4	45.7	3.097×10^{-6}	9.400×10^4	45.8	3.421×10^{-6}
8.850×10^4	46.1	3.268×10^{-6}	8.850×10^4	46.2	3.651×10^{-6}
8.300×10^4	46.4	3.507×10^{-6}	8.300×10^4	46.4	3.884×10^{-6}
7.750×10^4	46.6	3.767×10^{-6}	7.750×10^4	46.6	4.130×10^{-6}
7.200×10^4	46.7	4.077×10^{-6}	7.200×10^4	46.7	4.399×10^{-6}
6.650×10^4	46.8	4.430×10^{-6}	6.650×10^4	46.8	4.699×10^{-6}
6.100×10^4	46.9	4.787×10^{-6}	6.100×10^4	46.9	5.021×10^{-6}
5.550×10^4	47.0	5.221×10^{-6}	5.550×10^4	47.0	5.399×10^{-6}
5.000×10^4	48.2	5.794×10^{-6}	5.000×10^4	48.4	5.849×10^{-6}
4.552×10^4	49.1	6.392×10^{-6}	4.552×10^4	49.3	6.404×10^{-6}
4.108×10^4	49.5	7.057×10^{-6}	4.108×10^4	49.6	7.052×10^{-6}
3.664×10^4	49.8	7.889×10^{-6}	3.664×10^4	49.8	8.042×10^{-6}
3.220×10^4	50.0	9.005×10^{-6}	3.220×10^4	50.0	9.134×10^{-6}
1.000×10^3	57.9	2.886×10^{-4}	1.000×10^3	60.1	2.887×10^{-4}
5.000×10^2	65.4	5.773×10^{-4}	5.000×10^2	66.4	5.772×10^{-4}
1.000×10^2	75.0	2.886×10^{-3}	1.000×10^2	75.0	2.887×10^{-3}

ABSTRACT

Molecular dynamics is a well-known technique very much used in the study of biomolecular systems. The trajectory files produced by molecular dynamics simulations are extensive and the classical lossless algorithms give poor efficiencies in their compression. In this work, a new specific algorithm, named Byte Structure Variable Length Coding (BS-VLC), is introduced. Trajectory files, obtained by molecular dynamics applied to trypsin and trypsin:PTI complex, were compressed using four classical lossless algorithms (Huffman, adaptive Huffman, LZW and LZ77) as well as the BS-VLC algorithm. The results obtained show that BS-VLC nearly triplicates the compression efficiency of the best classical lossless algorithm, preserving a near lossless behavior. Compression efficiencies close to 50% can be obtained with a high degree of precision and the maximum efficiency possible (75%), within this algorithm, can be performed with good precision.

Role of Nitrogen Oxides in Ozone Toxicity

M. Friedman[1], S. Kazazić[2], N. Kezele[2], L. Klasinc[2,3], S. P. McGlynn[3], S. Pečur[2] and W. A. Pryor[3,4]

[1] Tulane University Medical Center, New Orleans, LA 70148, USA
[2] Ruđer Bošković Institute, HR 10 001 Zagreb, Croatia
[3] Chemistry Department, Louisiana State University, Baton Rouge, LA 70803, USA
[4] Biodynamics Institute, Louisiana State University, Baton Rouge, LA 70803, USA

Abstract

The preparation of ozone/nitrogen oxides mixtures in air containing the nitrate radical, their reaction with the unsaturated lipid 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine (POPC), and the determination of the reaction products in comparison to those obtained from a reaction with only ozone in air is described by MALDI-FTMS. The results indicate the importance of nitrate radical in ozone toxicity.

Introduction

Ozone is the most abundant oxidant in polluted air and its adverse effects on human health are well documented.[1] The lung is the organ that is most affected by ozone. Short term exposure to high levels of ozone leads to acute inflammatory reactions and pulmonary edema whereas prolonged exposure to lower levels produces emphysema, bronchopneumonia and fibrosis. Numerous studies have established the ability of ozone to react with species present in the lung; they include the amino acids, peptides and proteins.[2,3] However, the primary target of ozone is thought to be the unsaturated fatty acids (UFA) in the fluid layer of the lung

1

lining and in the epithelial cell membranes of the lung.[4-6] Because of its reactivity, ozone does not penetrate far into the cells that line the airways; consequently, many pulmonary and all extrapulmonary effects of ozone must be caused by messenger species. Thus, inhalation of even low ozone concentrations can cause the release of proinflammatory mediators in the lung, and it is these mediators that lead to the inflammation and other effects associated with ozone.

The cascade hypothesis[7] states that lipid ozonation products (LOP) relay the effects of ozone into deeper tissue strata at the lung-air interface than ozone itself can reach. LOP, rather than products from ozonation of proteins or nucleic acids, are thought to be signal transduction species because ozonation of UFA leads to small, diffusible, stable or metastable species, and because lipid oxidation products are known to act as signal transduction agents in other systems. Thus, the likely candidates for signal transduction species are LOP produced in the Criegee ozonation process, which gives a predictable spectrum of products.[8-11] Recent results by Friedman, Pryor and coworkers strongly support this hypothesis.[12-17]

Nitrogen oxides (NO and NO_2) are usually present in a mixture with other air pollutants in real-life exposures.[18,19] With ozone present, all nitric oxide is converted to nitrogen dioxide. Early studies indicated that the responses to ozone and nitrogen dioxide were additive, but it was also found that the immediate effects in rat lungs were dissimilar with respect to lipid peroxidation, lung protein or nonprotein sulfhydryl levels.[20] Starting in the 1980's, Pryor and coworkers[21-23] investigated the oxidation of biological molecules by nitrogen dioxide. The relatively high tolerance for both long-term and short-term exposure to ambient nitrogen dioxide made it unnecessary to include NO_2 chemistry in the cascade hypothesis of ozone toxicity. However, the combined action of ozone and nitrogen dioxide must take account of the fact that these gases rapidly react to form the nitrate radical, a very

2

potent oxidant. In fact, NO_3 reacts with unsaturated organic molecules more than a thousand times faster than does ozone.[24]

Here we report the preparation of ozone/nitrogen oxides mixtures in air containing the nitrate radical. The reaction of such mixtures is carried out with 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine (POPC) applied on a surface and the reaction products are determined by matrix-assisted laser desorption/ionization Fourier transform mass spectrometry (MALDI-FTMS). The results are compared with those obtained from parallel reaction when only ozone in air was used to react with POPC. The optimal conditions for determination the reaction products and elucidation of the synergistic effects of ozone and nitrate radical in the heterogeneous reaction with lipids are determined.

Experimental

Reactions. An apparatus (Fig. 2) consisting of an evacuable glass reaction column and mixing chamber, maintained in the dark because of the ready photodegradability of nitrate radical, was used to perform the reactions. The lipid, POPC (Sigma Chemical Co., St Louis, MO), dissolved in dichloromethane and spread inside the reaction column, yielded a thin film after evaporation. The lipid film was allowed to react with the gas mixture from the mixing chamber for a specified time. If necessary to ensure completion of the reaction, the previously reacted thin layer of sample was redissolved and, using the same procedure, was again exposed to the same gas mixture. The gas reaction mixtures were prepared by mixing streams of known concentrations of ozone in pure oxygen with a stream of pure nitrogen, which did or did not contain a known concentration of nitric oxide. The concentrations within the gas streams produced air samples with environmentally-relevant concentrations of ozone and nitrogen dioxide (appr. 100 ppb of each). Thus, in conditions of excess ozone, fast conversion

of nitric oxide produced O_3/NO_2 mixtures which, in the dark, react slower to give nitrate radicals. Either ozone/air or (ozone + nitrogen oxides)/air mixtures were allowed to react with the lipid films, and the reaction products of each reaction were compared.

Mass spectra. Mass spectra were recorded on an FTMS 2001 DD spectrometer (Finnigan FT/MS, Madison, WI, USA) equipped with a 3 T superconducting magnet, and a Nicolet 1280 data station using a pulsed nitrogen laser (VSL-337ND-S, Laser Science, Inc. Franklin, MA, USA) at 337 nm for MALDI experiments. We initially used 2,5-dihydrobenzoic acid (DHB) as the MALDI matrix. It gave abundant positive and negative fragmentation ions for the POPC and its ozonation products but little or no protonated (m/z = 760), sodiated (m/z = 782) and ozonated (m/z = 808) molecular ions. To avoid fragmentation we tested unsuccessfully several other matrices (e.g. *p*-nitrobenzoic acid and 3-nitrophenol) and sample probe cooling. Only 3-nitrobenzyl alcohol as a matrix with a cooled sample probe yielded phospholipid ions and mass spectra of their reaction products with little or no fragmentation and formation of alkali metal adduct peaks. All the spectra were collected after a single laser shot.

Calculations. Because the rate constants for all the respective reactions are known (Table 1) it is possible to calculate development of NO_3 radical concentration starting with O_3 + NO at room temperature and normal pressure in air. Calculation results for the first 1000 s of 50 ppb NO with 100, 150 and 200 ppb of ozone show final concentrations of 0.11, 0.24 and 0.60 ppb for NO_3, and 48, 96 and 141 ppb for ozone, respectively. In these calculations the photoreaction of NO_3 (dark conditions) and decomposition of ozone (amounting less than 10 % for the investigated time period as independently determined) were omitted. The results are shown in Fig. 2a-c.

The ratio of O_3/NO_3 in our reaction conditions is expected to be ~ 500 which gives the nitrate radical a more than twentyfold advantage over ozone in reaction with POPC. Although N_2O_5 is present in higher concentration than is NO_3, it is of only minor importance because it reacts with water yielding HNO_3 which will remain bound.[25]

Results and Discussion

Compared to prior ionization techniques, phospholipid analysis with MALDI-FTMS provides much higher mass resolving power and sensitivity.[26-40] The only studies of lipid ozonation products but not with NO_3 present were done by Finlayson-Pitts and coworkers[38,39] using fast atom bombardment (FAB) and the investigation of phospholipids by Marshall and coworkers[40] using MALDI-FTMS. The latter study is very useful for the present investigation because it provides an experimental fragmentation scheme for POPC and introduces the use of cooled matrices which is crucial for the study of reaction products with ozone and nitrate radical.

Using a 3-nitrobenzyl alcohol matrix (~ 3000:1 matrix-to-analyte ratio) we observe negligible fragmentation of the lipid and a weaker sodiated molecular ion which sometimes is not observed. The mass spectra of POPC and its reaction products either with ozone or with ozone containing NO_3 obtained using a 3-nitrobenzyl alcohol matrix are shown in Fig. 3a-c. The reaction of POPC with ozone is expected to proceed via the unstable primary ozonide which fragments to the secondary (Criegee) ozonide via zwitterionic species (Scheme 1). The Criegee ozonide decomposes yielding an acid and aldehyde pair; i.e., either PC/AC + ALD/C9 or PC/ALD + AC/C9, respectively (see Scheme 1 for definitions and structure). The MALDI mass spectrum of the products of incomplete ozonation reaction of POPC (Fig. 3b), which is similar to that obtained by using FAB[38], confirms the products predicted by

Scheme 1 indicating a slightly higher probability for AC/C9 than ALD/C9 formation (i.e. higher PC/ALD than PC/AC peak).

The comparison of the products after extensive reaction of POPC (Fig. 3c) with ozone (LHS) and the ozone/nitrogen oxides mixture (RHS) shows the products are similar but with a dramatic change in their yields. In the mass spectrum of reaction products with ozone, a new ion of $m/z = 638$ is observed which could correspond to loss of O_2 from some peroxide type product structure ($X-O_2$) whereas in the reaction products including nitrate radical an unknown ion of $m/z = 623$ appears which could correspond to loss of NO_2 from a similar nitrite structure (X-ONO). The formation of new products with ($O_3 + NO_x$) in dark is not surprising since NO_3 exhibits much higher reactivity with unsaturated organics than ozone.[24]

These results indicate the importance of using realistic and more complex mixtures of oxidants in study of ambient ozone toxicity. Clearly, further study of these complex mixtures will be rewarding.

Acknowledgement

References

(1) Lippman, M. Health effects of ozone. *J. Air Pollut. Contr. Assoc.* **1989**, 39, 672-695.

(2) Pryor, W. A.; Dooley, D. F. and Church, D. F. Mechanisms for the reaction of ozone with biological molecules: The source of the toxic effects of ozone. In: *Advances in Modern Environmental Toxicology, Volume V. The Biomedical Effects of Ozone and Related Photochomical Oxidants* (Lee, S. D., Mustafa, M. G. and Mehlman, M. A. eds.) Princeton Scientific Publishers, Princeton, NJ, 1982, pp 7-19.

(3) Uppu, R. M. and Pryor, W. A. The reactions of ozone with proteins and unsaturated fatty acids in reverse micelles. *Chem. Res. Toxicol.* **1994**, 7, 47-55.

(4) Postlethwait,E.M.; Cueto, R.; Velsor, L.W. and Pryor, W.A. O_3-induced formation of bioactive lipids: estimated surface concentrations and lining layer effects. *Am. J. Physiol.* **1998**, 274 (*Lung Cell. Mol. Physiol.* 18), L1006-L1016.

(5) Kafoury, R.M.; Pryor, W.A.; Squadrito, G.L.; Salgo, M.G.; Zou, X. and Friedman, M. Lipid ozonation products activate phospholipase A2, C, and D. *Toxicol. Appl. Pharmacol.* **1998**, 150, 338-349.

(6) Leikauf, G. D.; Zhao, Q.; Zhou, S. and Santrock, J. Ozonolysis products of membrane fatty acids activate eicosanoid metabolism in human airway epithelial cells. *Am. J. Respir. Cell Mol. Biol.* **1993**, 9, 594-602.

(7) Pryor, W. A.; Squadrito, G. L. and Friedman, M. The cascade mechanism to explain ozone toxicity: the role of lipid ozonation products. *Free Radic. Biol. Med.* **1995**, 19, 935-941.

(8) Pryor, W. A. and Wu, M. Ozonation of methyl oleate in hexane, in a thin film, in SDS micelles, and in distearoylphosphatidylcholine liposomes: yields and properties of the Criegee ozonide. *Chem. Res. Toxicol.* **1992**, 5, 505-511.

(9) Squadrito, G. L.; Uppu, R. M.; Cueto, R. and Pryor, W. A.() Production of the Criegee ozonide during the ozonation of 1-palmitoyl-2-oleoyl-syn-glycero-3-phosphocholine liposomes. *Lipids* **1992**, 27, 955-958.

(10) Criegee, R. In: *Peroxide Reaction Mechanisms Edwards*, J. D.; Ed.; Wiley-Interscience: New York,1962

(11) Bailey, P. S. *Ozonation in Organic Chemistry*, Vol 1, Academic Press: New York, 1978

(12) Pryor, W. A.; Das, B. and Church, D. F. The ozonation of unsaturated fatty acids: aldehydes and hydrogen peroxide as products and possible mediators of ozone toxicity. *Chem. Res. Toxicol.* **1991**, 4, 341-348.

(13) Cueto, R.; Squadrito, G. L. and Pryor, W. A. Quantifying aldehydes and distinguishing aldehydic product profiles from autoxidation of unsaturated fatty acids. *Methods Enzymol.* **1994**, 233, 174-182.

(14) Pryor, W. A. and Church, D. F. The reaction of ozone with unsaturated fatty acids: aldehydes and hydrogen peroxide as mediators of ozone toxicity. In: *Oxidative Damage & Repair: Chemical, Biological and Medical Aspects*, pp 496-504, Pergamon Press, New York, 1991

(15) Church, D. F.; McAdams, M. and Pryor, W. A. Free radical production from the ozonation of simple alkenes, fatty acid emulsions and phosphatidylcholine liposomes. In: *Oxidative Damage & Repair: Chemical, Biological and Medical Aspects*, Pergamon Press, New York, 1991, pp 517-522.

(16) Pryor, W. A.; Stanley, J. P.; Blair, E. and Cullen, G. B. Autoxidation of polyunsaturated fatty acids. Part I. Effect of ozone on the autoxidation of neat methyl linoleate and methyl linolenate. *Arch. Environ. Health* **1976**, **31**, 201-210.

(17) Pryor, W. A.; Stanley, J. P. and Blair, E. Autoxidation of polyunsaturated fatty acids. Part II. A suggested mechanism for the formation of TBA-reactive materials from prostaglandin-like endoperoxides. *Lipids* **1976**, 11, 370-379.

(18) Penkett, S. A.; Blake, N. J.; Lightman, P.; Marsh, A. R. W.; Anwyl, P. and Butcher, G. The seasonal variation of nonmethane hydrocarbons in the free troposphere over the north Atlantic ocean: Possible evidence for extensive reaction of hydrocarbons with the nitrate radical. *J. Geophys. Res.* **1993**, 98, 2865-2885.

(19) Aschmann, S. M. and Atkinson R. Rate constants for the reactions of the NO_3 radical with alkanes at 296 ± 2 K. *Atmos. Environ.* **1995**, 29, 2311-2316.

(20) Lindvall, T. Health effects of nitrogen dioxide and oxidants. *J. Work Environ. Health* **1985**, 11 (Suppl. 3), 10-28.

(21) Pryor, W. A.; Prier, D. G.; Lightsey, J. W. and Church, D. F. Initiation of the autoxidation of polyunsaturated fatty acids (PUFA) by ozone and nitrogen dioxide. In: *Autoxidation in Food and Biological Systems* (Simic, M. G. and Karel, M. eds.), Plenum Publishing Corp., New York, 1980, pp 1-16.

(22) Pryor, W. A.; Lightsey, J. W. Mechanisms of nitrogen dioxide reactions: Initiation of lipid peroxidation and the production. *Science* **1981**, 214, 435-437.

(23) Gallon, A. A. and Pryor, W. A. The reaction of low levels of nitrogen dioxide with methyl linoleate in the presence and absence of oxygen. *Lipids* **1993**, 29, 171-176.

(24) Atkinson, R.; Arey, J.; Aschmann, S. M.; Corchnoy, S. B. and Shu, Y. Rate constants for the gas-phase reactions of cis-3-hexen-1-ol, cis-3-hexenylacetate, trans-2-hexanal, and

linalool with OH and NO$_3$ radicals and O$_3$ at 296 ± 2 K, and OH radical formation yields from the O$_3$ reactions. *Int. J. Chem. Kinet.* **1995**, 27, 941-955

(25) Schurath, U., personal communication.

(26) Klein, R. A. Mass spectrometry of the phosphatidylcholines: dipalmitoyl, dioleoyl, and stearoyl-oleoyl glyceryl phosphorylcholines. *J. Lipid Res.* **1971**, 12, 123-131.

(27) Klein, R. A. Mass spectrometry of the phosphatidylcholines. Fragmentation processes for dioleoyl and stearoyl-oleoyl glyceryl-phosphorylcholine. *J. Lipid Res.* **1971**, 12, 628-634.

(28) Wood, G. W. and Lau, P. Y. Analysis of intact phospholipids by field desorption mass spectrometry. *Biomed. Mass Spectrom.* **1974**, 1, 154-155.

(29) Wood, G. W.; Lau, P. Y. and Rao, G. N. S. () Field desorption mass spectrometry of phospholipids: Fragmentation of dipalmitoylphosphatidylcholine from comparision of d0, d4 and d9 species. *Biomed. Mass Spectrom.* **1976**, 3, 172-176.

(30) Wood, G. W.; Lau, P. Y.; Morrow, G.; Rao, G. N. S. and Schmidt, D. E. Field desorption mass spectrometry of phospholipids: Survey and structural types. *J. Chem. Phys. Lipids* **1977**, 18, 316-333.

(31) Crawford, C. G. and Plattne, R. D. Ammonia chemical ionization mass spectrometry of intact diacyl phosphatidylcholine. *J. Lipid Res.* **1983**, 24, 456-460.

(32) Bisseret, P.; Nakatani, Y.; Ourisson, G.; Hueber, R. and Teller, G. Ammonia chemical ionization mass spectrometry of lecithins on a gold support *J. Chem. Phys. Lipids* **1983**, 33, 383-392.

(33) Demirev, P. A. Californium-252 plasma desorption mass spectrometry of glycerophospholipids. *Biomed. Mass Spectrom.* **1987**, 14, 241-246.

(34) Cotter, J. R. and Tabet, J. C. Laser desorption mass spectrometry: mechanisms and applications. *Int. J. Mass Spectrom. Ion Phys.* **1983**, 53, 151-166.

(35) Fenwick, G. R.; Eagles, J. and Self, R. Mass-analyzed ion kinetic energy spectra and B1E-B2 triple sector mass spectrometric analysis of phosphoinositides by fast atom bombardment. *Biomed. Mass Spectrom.* **1983**, 10, 382-386.

(36) Jensen, N. J.; Tomer, K. B. and Gross, M. L. Fast atom bombardment and tandem mass spectrometry of phosphatidylserine and phosphatidylcholine. *Lipids* **1986**, 21, 580-588

(37) Jensen, N. J.; Tomer, K. B. and Gross, M. L. FAB MS/MS for phosphatidylinositol, -glycerol, -etanolamine and other complex phospholypids. *Lipids* **1987**, 22, 480-489

(38) Lai, C. C.; Finlayson-Pitts, B. J. and Willis, W. A. Formation of secondary ozonides from the reaction of an unsaturated phosphatidylcholine with ozone. *Chem. Res. Toxicol.* **1990**, 3, 517-523.

(39) Finlayson-Pitts, B. J.; Pham, T. T. H.; Lai, C. C.; Johnson, S. N.; Luciogough, L. L.; Mestats, J. and Iwig D. Thermal decomposition of phospholipid secondary ozonides – implications for the toxicity of inhaled ozone. *Inhalation Toxicology.* **1998**, 10, 813-830.

(40) Marto, J. A.; White, F. M.; Seldomridge, S. and Marshall, A. G. Structural characterization of phospholipids by matrix-assisted laser desorption/ionization Fourier transform ion cyclotron resonance mass spectrometry, *Anal. Chem.* **1995**, 67, 3979-3984.

11

Table 1. Reactions and their rate constants used in the simulation of product formation from ozone/nitric oxide mixtures in air.

Reaction	Rate constants[*]
$NO + O_3 \rightarrow NO_2 + O_2$	1.8×10^{-14} $cm^3molecule^{-1}s^{-1}$
$NO + NO_3 \rightarrow 2NO_2$	2.6×10^{-11} $cm^3molecule^{-1}s^{-1}$
$NO_2 + O_3 \rightarrow NO_3 + O_2$	3.2×10^{-17} $cm^3molecule^{-1}s^{-1}$
$NO_2 + NO_3 + M \rightarrow N_2O_5$	2.0×10^{-12} $cm^3molecule^{-1}s^{-1}$
$N_2O_5 + M \rightarrow NO_2 + NO_3 + M$	6.9×10^{-2} s^{-1}
$NO_3 + NO_3 \rightarrow 2NO_2 + O_2$	8.5×10^{-13} $cm^3molecule^{-1}s^{-1}$

* from CRC Handbook of Chemistry and Physics, 79th Edition, David R. Lide, ed., CRC Press, Boca Raton, 1998, p. 5-105.

$$CH_2O. \overset{\overset{O}{\|}}{C}. (CH_2)_{14}.CH_3$$

$$CHO. \overset{\overset{O}{\|}}{C}. (CH_2)_7CH=CH(CH_2)_7CH_3 \quad + \quad O_3$$

$$CH_2O. \overset{\overset{O}{\|}}{P}.O.(CH_2)_2.N^+(CH_3)_3$$
$$\overset{\downarrow}{O^-}$$

⇓

$$CH_2O. \overset{\overset{O}{\|}}{C}. (same)$$

$$CHO. \overset{\overset{O}{\|}}{C}. (CH_2)_7HC—CH.(CH_2)_7CH_3 \qquad \text{PRIMARY OZONIDE}$$

$$CH_2O. (same)$$

$$CH_2O. \overset{\overset{O}{\|}}{C}. (CH_2)_{14}.CH_3$$

$$CHO. \overset{\overset{O}{\|}}{C}. (CH_2)_7.CHO \quad + \quad {}^-OOC^+ \overset{H}{\underset{(CH_2)_7CH_3}{}}$$

$$CH_2O. \overset{\overset{O}{\|}}{P}.O.(CH_2)_2.N^+(CH_3)_3$$
$$\overset{\downarrow}{O^-}$$

$$CH_2O. \overset{\overset{O}{\|}}{C}. (CH_2)_{14}.CH_3$$

$$CHO. \overset{\overset{O}{\|}}{C}. (CH_2)_7HC^+. OO^- \quad + \quad OCH \overset{}{\underset{(CH_2)_7}{\underset{CH_3}{}}}$$

$$CH_2O. \overset{\overset{O}{\|}}{P}.O.(CH_2)_2.N^+(CH_3)_3$$
$$\overset{\downarrow}{O^-}$$

$$CH_2O. \overset{\overset{O}{\|}}{C}. (same)$$

$$CHO. \overset{\overset{O}{\|}}{C}. (CH_2)_7HC \underset{a}{\overset{}{}} O \underset{b}{\overset{}{}} CH(CH_2)_7CH_3 \qquad \text{CRIEGEE (SECONDARY) OZONIDE}$$

$$CH_2O. (same)$$

a

b

$$CH_2O. \overset{\overset{O}{\|}}{C}. (CH_2)_{14}.CH_3$$

$$CHO. \overset{\overset{O}{\|}}{C}. (CH_2)_7.CHO \quad + \quad COOH \overset{}{\underset{(CH_2)_7}{\underset{CH_3}{}}}$$

$$CH_2O. \overset{\overset{O}{\|}}{P}.O.(CH_2)_2.N^+(CH_3)_3$$
$$\overset{\downarrow}{O^-}$$

$$CH_2O. \overset{\overset{O}{\|}}{C}. (CH_2)_{14}.CH_3$$

$$CHO. \overset{\overset{O}{\|}}{C}. (CH_2)_7COOH \quad + \quad OCH \overset{}{\underset{(CH_2)_7}{\underset{CH_3}{}}}$$

$$CH_2O. \overset{\overset{O}{\|}}{P}.O.(CH_2)_2.N^+(CH_3)_3$$
$$\overset{\downarrow}{O^-}$$

PC/ALD   **AC/C9**   **PC/AC**   **ALD/C9**

Scheme 1. Product formation in the reaction of POPC with ozone according to the Criegee ozonation process.
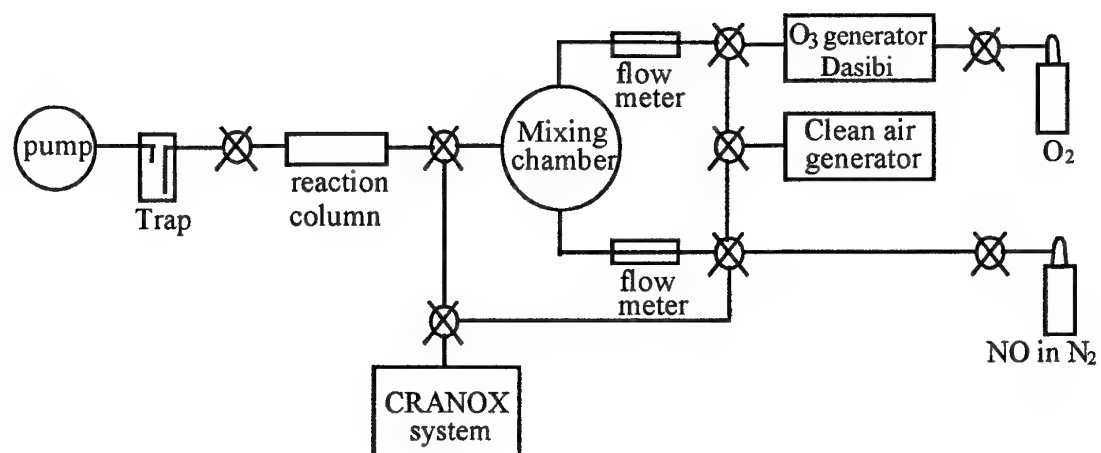
Figure 1. Apparatus for carying out the reaction of POPC with ozone and ozone/nitric oxide mixtures in air.
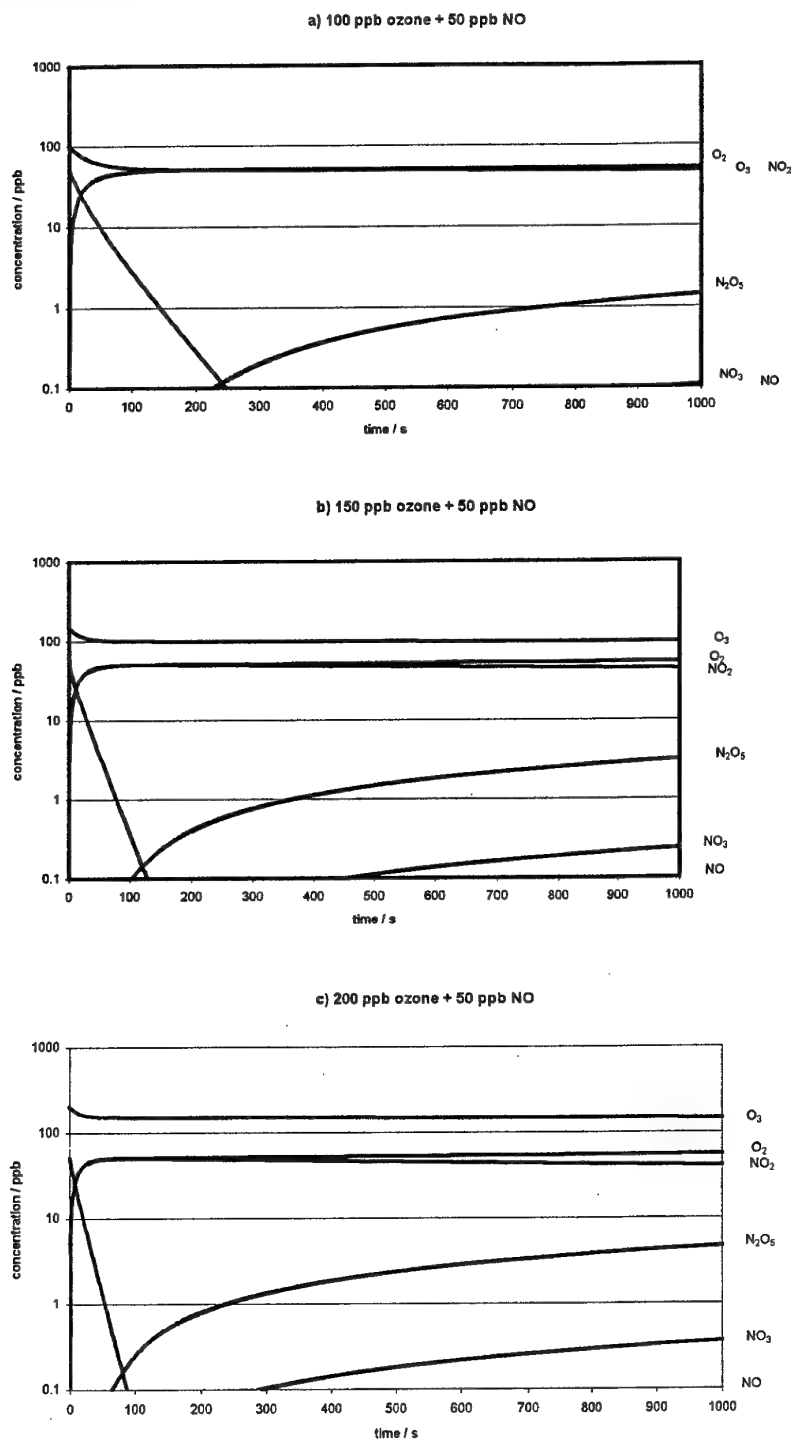
Figure 2. Simulation of first thousand second period of product development from a mixture of 50 ppb of NO with
a) 100 ppb
b) 150 ppb and
c) 200 ppb of ozone in air using rate constants from Table 1.

(a)   **FTMS OF POPC**
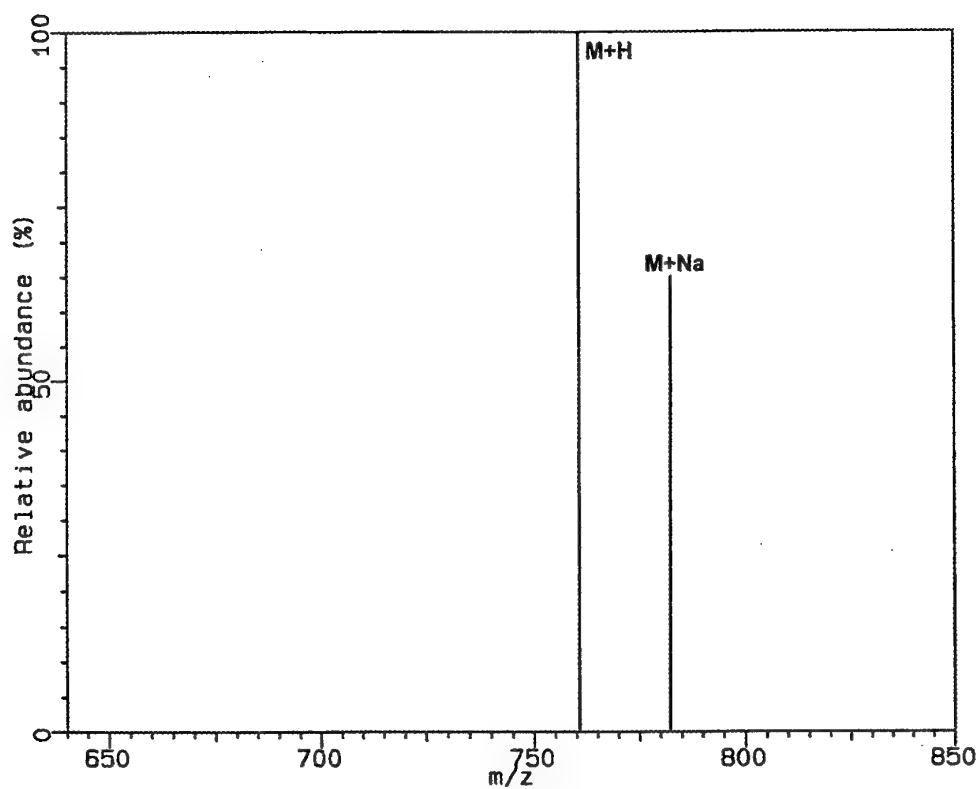
(b)   **FTMS OF POPC AFTER 10 min REACTION WITH O$_3$**
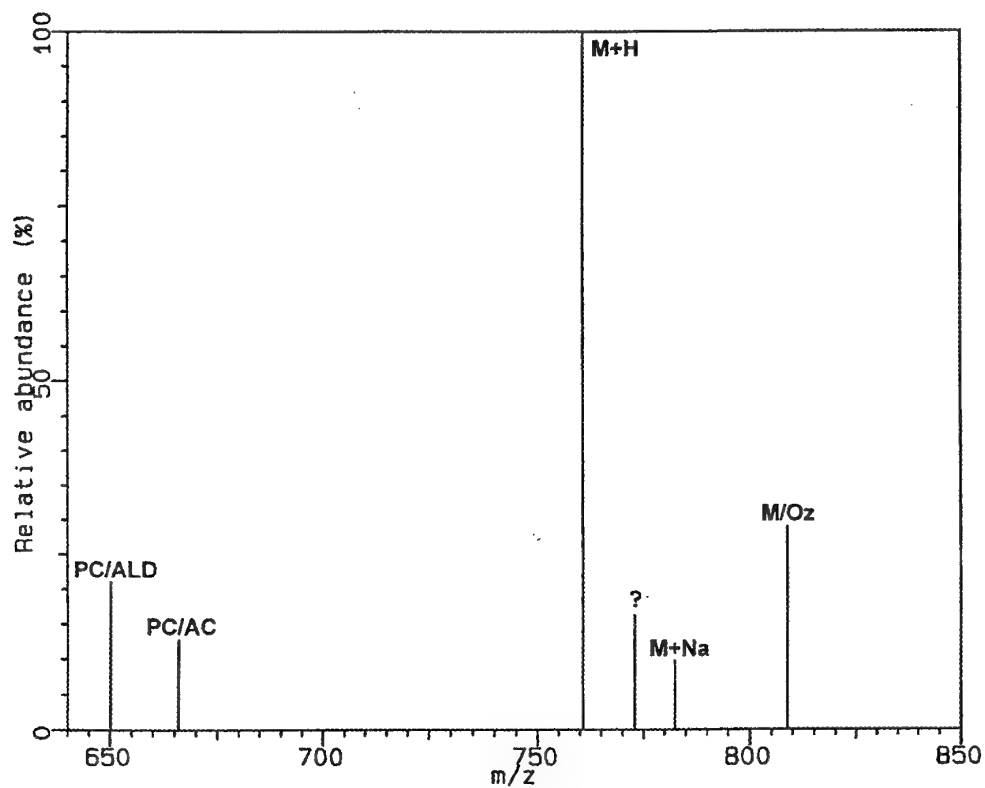
(c)
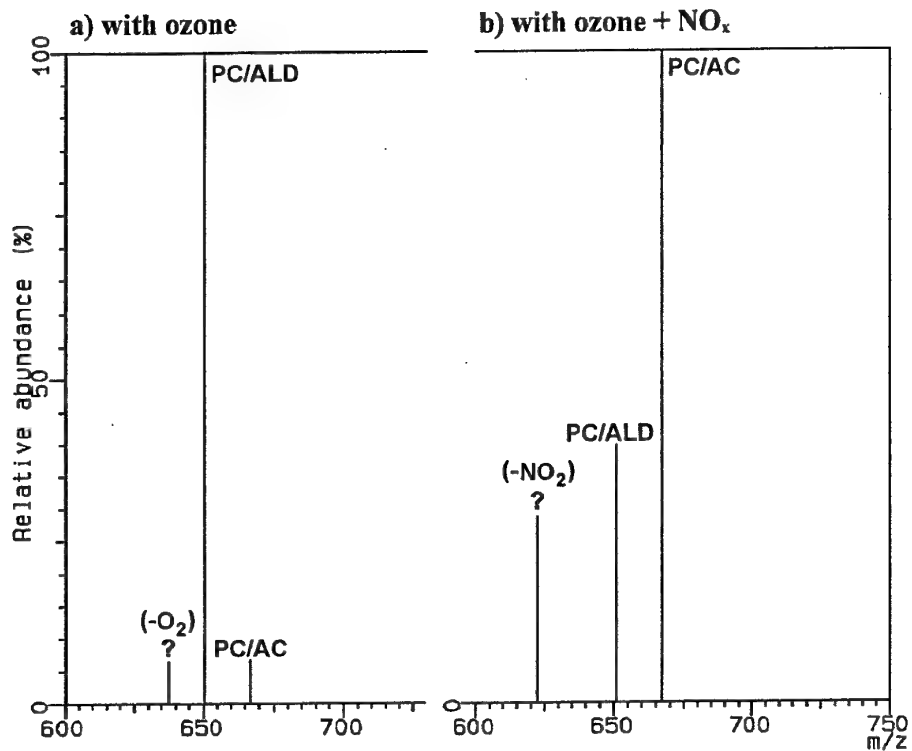
Figure 3: MALDI FTMS positive ion spectra from single nitrogen laser shots on cryogenic

matrices of 3-nitrobenzyl alcohol in 3000:1 ratio with the following analytes:

a) pure POPC

b) POPC after 10 min exposure to ozone and

c) reaction products of POPC with ozone (LHS) and ozone/nitrogen oxides mixture (RHS).

17

UNIVERSAL METRIC PROPERTIES OF THE GENETIC CODE

Nikola Štambuk

Rudjer Bošković Institute, Bijenička 54, HR-10001 Zagreb, Croatia

Correspondence to: Dr. Nikola Štambuk, Šubićeva 16, HR-10000 Zagreb, Croatia.

e-mail: stambuk@rudjer.irb.hr

ABSTRACT

Universal metric properties of the genetic code (i.e. RNA, DNA and protein coding) are defined by means of the nucleotide base representation on the square with vertices U or T = 00, C = 01, G = 10 and A = 11. It is shown that this notation defines Cantor set and Smale horseshoe map representation of the genetic code, classic table arrangement and Siemion one-step mutation ring of the code. Gray code solution to the problem with all codon positions, and an extention to octal coding system are given. Finally, unified concept of the genetic code linked to the Cantor set and horseshoe map is introduced in the form of a classic combinatorial 4 colour necklace model with three horizontal frames of 64 coloured pearls (bases) and vertically hanging decorations of triplets (codons). Three horizontal necklace frames define Crick's code without comma, and vertical necklace decorations define the evolutional code. Thus, the type of the code depends on the level or direction of the observation. Fibonacci dynamics and Cantor set-Farey tree partition of codon and amino acid groups are discussed and explained. This method of genetic code analysis is named SCA procedure.

Key words: RNA, DNA, protein, genetic code, Gray code, circular, necklace, Cantor set, horseshoe map, golden mean, Farey tree

INTRODUCTION

The protein coding and synthesis in biological systems is, together with all other information of the genome, found in DNA and RNA strings consisting of 4 nucleotide base combinations (U or T, C, A and G).[1-3] The four bases define 64 codon triplets that specify 20 amino acids and 3 stop codons for the protein synthesis.[1-3] The aim of this paper is to define the universal metric properties of the codon and nucleotide base recombination. This will be done by addressing three dimensions of the problem, as follows.

First, we show that the quadratic binary representation of the 4 bases on the unit square may be projected on the Cantor set for all codons and amino acids. It is proved that for the one-dimensional projection symbolic binary coordinates provide the Gray code solution to the problem of amino acid coding. Counter-clockwise and clockwise changes of the base positions on the square define the link of the classic genetic code table and Siemion one-step mutation ring of the genetic code (which is linked to the physico-chemical properties of the amino acids).

Second, we show that Smale horseshoe map representation of binary and Cantor codon (amino acid) positions defines classic table of the genetic code. This result shows that the genetic code table is a reflection of the standard horseshoe map, often used in the analysis of nonlinear and chaotic systems. The possibility of analysis by binary and octal coding system is discussed and octal code addresses for codons and amino acids are given.

Third, we show that a classic combinatorial 4 colour necklace problem[4], with each colour representing a nucleotide base projection on the unit square, defines the unified concept of the

genetic code. Three horizontal frames of the necklace, consisting of 64 coloured pearls (bases), make Crick's comma-less code and vertically hanging decoration triplets (codons) define evolutional code. Thus, it is proved that the necklace model defines both concepts, depending on the level of observation and/or position of the observer.

Finally, the process of base, codon and amino acid recombination in the genetic code table is discussed with respect to their Cantor set and Farey tree partition frequencies. It is emphasised that this interaction leads to the Fibonacci, i.e. golden ratio, based dynamics in selecting codon and amino acid families, as previously discussed by Schroeder and Štambuk.[5,6] This method of genetic code notation and analysis is named SCA procedure.

RESULTS AND DISCUSSION

Primary coding problem - metric on the unit interval

The notation

We introduce the binary representation of 4 nucleotide bases on the square with vertices 00, 01, 10, 11 in a manner defined for Cantor set by H. Steinhaus in 1917 (when discussing interesting properties of the set noticed by S. Banach).[7] The notation U or T = 00, C = 01, G = 10 and A = 11 is presented in Figure 1. It has the following properties:

The combination of 2 digits (0 or 1), denoting *primary* and *secondary characteristics* of the nucleotide base describe each of the letters according to the group subdivision/discrimination principles (1st digit purine-pyrimidine, 2nd digit strong-weak H bond discrimination). The first and weak H bonding pyrimidine base U or T = 00 is discriminated from the next strong H bonding pyrimidine base C = 01 by the second digit notation. Full complementarity in obtaining weak (A)

and strong H bonding purines (G) is achieved by $0 \leftrightarrow 1$ pyrimidine changes (to A = 11, G = 10), or vice versa.

Codon positions on the binary tree

Table 1 shows the binary notation for all 64 codons and 20 amino acids. To define more precisely the positions of particular codon intervals of the *binary tree* with respect to the quadratic base mapping we examine invariant Cantor set with the method of symbolic dynamics in a standard manner.[3,8] This was performed since Cantor set possesses two properties related to the binary coding of the Figure 1 notation:[3,8]

1. binary decomposition of the initial segment into 2^n segments projected on $(n-1)^{th}$ binary tree level,

2. partitioning of the observed set by excluding 1/3 of its original length per each of the tree levels.

The relative location of different coding intervals and their orientation is additionally specified in Table 1, by the nodes of *alternating binary tree* and their symbolic coordinates (names).[8] Briefly, the left half of the unit interval is labelled 0 and the right one 1. For $x<1/2$ and its derivative $f_\lambda(x)>0$, with $f(x) = \lambda x(1-x)$, $\lambda \geq 4$, the pairs of the initial binary tree preserve orientation and for $x>1/2$, $f_\lambda(x)<0$ they reverse orientation in the alternating binary tree.[8]

Gray code solution to the metric problem

Symbolic coordinates of codon and amino acid locations on the Cantor set in Table 1 represent the Gray code solution to the n=6 digit binary notation for $2^n = 64$ codons. This result has been published by M. Gardner in 1972.[9] Gardner's Gray numbers that solve the puzzle for n=6 digits/rings are symbolic addresses of different codons in Table 1, and Cantor set solution to this problem represents their projection to [0, 1] interval according to their appearance. Consequently, the stretching and folding of the quadratic map with symbolic dynamics on the unit interval,[8] keeps track and

information of the hypercube codon (amino acid) representations by means of the Gray code. Two dimensional representation is defined via horseshoe map.

The unit interval Cantor mapping in Table 1 solves complementary coding problem via binary tree codon projection, since Gray code solution requires at least 32 binary numbers from the first part of the table. Complementary addresses for the second half of the table are symmetrically arranged at opposite Cantor positions and obtained by $0 \leftrightarrow 1$ digit switch.

Siemion's mutation ring and genetic code table

Table 2 shows that permutations of the 4 amino acid families in classic genetic code Table 3 (CUGA, UGAC, GACU and ACUG) identify Siemion's one-step mutation ring of the genetic code[10] presented in Table 4. This is done by means of UC/CU, AG/GA replacements (U row/column), C/G - G/C mutation (C row/column), A/U mutation (A row/column) and C/G mutation (G row/column). Four amino acid families are defined by a simple algorithm of the binary codon notation (Figure 1).

The extraction of the Siemion's mutation ring from the standard genetic code table, by means of the Cantor set based nucleotide notation (and algorithm) in Figure 1, is of considerable importance since Siemion's ring is related to the physico-chemical properties of different amino acids.[3,10]

Secondary coding problem - Horseshoe map metric and octal coding

Smale's Horseshoe Map

The Smale horseshoe map is the example of a chaotic hyperbolic invariant set and the map often behaves like skeleton on which chaotic and periodic orbits of the system are organized.[8,11] The horseshoe is a mapping of the unit square (Figure 1) which contracts the horizontal directions, expands in the vertical direction, and then folds. The mapping is only defined on the unit square and

points that leave the square are ignored.[8] Forward and backward iteration of the horseshoe map generate the locations of the periodic points.[8,11]

Amino acid and codon horseshoe mapping

By iterating the map we specified the locations of a periodic orbits within the homoclinic tangle of the horseshoe. Table 5 gives the labelling scheme for horizontal and vertical branches from a pair of alternating binary trees. The projections of 2 binary triplets (or 2 octal numbers) according to the horseshoe pattern extract standard table of the genetic code (Table 3), which proves that this map defines the patterns of the codon recombination buried in the code. Patterns of the first, second and third base changes also satisfy and confirm standard square notation with 4 binary addresses presented in Figure 1, typical of the horseshoe map. The algorithm in Figure 1 is therefore confirmed for the genetic code and Table 5 represent its proper labelling scheme.

Since the invariant horseshoe set is a product of two Cantor sets intersections in horizontal and in vertical directions,[8] the Cantor set projection of the genetic code is also proved for a two-dimensional case.

Octal coding

Further extension of the coding system in Table 6 defines 28 pairs of all possible 8 (node) 3-dimensional cube permutations (i.e. as 8×8 codon octades).[3,12] Three binary digits define octal number coding, and $28 = 8!/2! (8-2)!$ pairing combinations are obtained from the permutations of 2 corresponding octal numbers[13] (or 2 binary triplets). Eight identical doublets in addition to 56 different dublets define all 64 codons. This pattern is consistent with three letter alphabet permutation consisting of two binary, i.e. 0 or 1, choices in the truth table $(2^3 = 8)$.[14]

Tertiary coding problem - Necklace model of the genetic code

<u>Circular code arithmetic and Necklace coding</u>

The genetic and protein circular code is defined by means of a combinatorial necklace model.[4] This structure consists of 64 beads of 4 different colours representing 4 nucleotide bases (U or T, C, A, G). The coloured beds are making decorations that consists of vertically hanging chains of $x = 3$ beds, which represent each of the codons. Consequently, there are $y = 4^3$ distinct vertical chains that can be made (i.e. number of words of length $x = 3$ with alphabet of size $y = 4$). The total number of possible vertical decorations containing at least two colours each is $y^x - y = 60$, and $y = 4$ decorations contain the beads of the same colour.

One of the characteristics of this system is that we may define "beheading" as the process where the top bead is taken and replaced on the bottom.[4] After some repetitions we observe the initial pattern. Let b be the smallest positive number of successive behadings (including reverse ones) needed to get back the original, we have:

$$1 < b \leq 3, \quad x = ab + c \ (0 \leq c < b).$$ (1)

The initial pattern is restored by x beheadings followed by a lots of b reverse behadings. For $c = 0$, $x = ab$, if x is prime and $b > 1$ we have $b = x, a = 1$. By observing the chains and their first $x - 1$ beheadings different collections are made (that cannot be transformed into each other). Thus, when $y^x - y$ chains have been accounted, we get a total of n collections $y^x - y = nx$ and $y^x = y \bmod x$, from which we obtain Fermat's theorem.[4]

<u>Coding patterns and codon collections</u>

Table 7.a-f presents the circular and complementary coding patterns for all possible codon collections. It contains two and three colouring collections consisting of 3 transformed/beheaded codons (12 and 8 collections of 3 triplets, i.e. 60, of 2 and 3 colours respectively) and four triplets of the same colour. It is shown that there exists the codon arrangement for each of 3 horizontal

necklace frames (m = mod 1, 2, 3) that is 100% identical to the empirically detected one by Arques and Michel,[1,2] Four triplets of the same colour link the endpoints of the frames enabling the construction of the three frame automaton (Table 8).

Frame shifts and frame retrieval

Table 8 shows that the arrangement of the codons in the frames according to their projection on the Cantor set, transforms each frame in such way that if one letter shift is performed the next frame is automatically retrieved (a-d). Few letter changes that occur during the transformations are permissible and predicted according to Molecular Recognition Theory (R⇔S, Q⇔H, D⇔E) or N-end rule (K⇔N),[15,16] i.e. the coding pattern is consistent with theoretical and empirical observations.

Unified concept of the genetic code

Presented results indicate that the concepts of code without comma or of evolutionary code, based on different premises, strongly depend on the level of the observation/analysis. In the necklace model Crick's code without comma[1,2,17] represents three horizontal frames that define necklace chains, while Dounce's evolutionary code[1,2,18] makes vertically hanging beds (codon triplets). Therefore circular coding necklace algorithm represents an unifying concept of the genetic code. This method, denoted SCA, enables the genetic code and protein analysis via number theory arithmetic for codes.

Concluding remarks

Fibonacci dynamics and Farey tree

Two dimensional Cantor set projection of the binary (square) notation via Smale horseshoe map reconstructs the classic table of the genetic code, which proves our result and opens the possibility for the gene and protein analyses as chaotic dynamical systems. Additionally, the closest intersections of Cantor set (binary & symbolic) and Farey tree codon projections define "golden amino acids" (related to the Fibonacci dynamics[19]). The Fibonacci dynamics, noticed in the algorithms of the

genetic code[3,10,19] and in long range DNA correlation exponents,[20] might arise from two mapping frequencies of the code. The frequency of Cantor set projection/recombination of the codons (2/3) is mixed with the frequency (1/2) of the Farey tree that splits the amino acid/codon groups upon each Cantor set projection level ($2^6 = 64$). As shown by Schroeder,[5] the frequency resulting from such Cantor set-Farey tree interaction is the golden ratio $0.618 = 2+1/3+2 = 3/5$, which explains previously mentioned phenomena. Some mathematical and dynamical aspects of those interactions have been discussed by Schroeder and Štambuk.[5, 6]

TG/CT excess, TA/CG deficiency and language decoding

Binary coding quadratic algorithm (Figure 1) based on the pyrimidine-purine (T-G) and strong-week H bonding (T-C) discrimination is in accordance with the universal rule of TG/CT excess and TA/CG deficiency in coding and noncoding DNA regions[21,22], since TA/CG does not satisfy second digit strong-week discrimination (and consequently may be less likely to appear). Another important aspect of this study is related to the discovery that non-coding DNA sequences posses properties characteristic of natural languages, while the coded DNA sequences correspond to the coded language structures.[23,24] In this context the concept presented in this study may contribute to the extraction/decoding of the programming language of DNA and RNA strings.

REFERENCES

1. Arques, D.G.; Michel, C.J. A complementary circular code in the protein coding genes. *J. Theor. Biol.* **1996**, 182 45-58.

2. Arques, D.G.; Fallot J.P.; Michel, C.J. An evolutionary model of complementary circular code. *J. Theor. Biol.* **1997**, 185, 241-253.

3. Štambuk, N. On the genetic origin of complementary protein coding. *Croat. Chem. Acta,* **1998**, 71, 573-589.

4. Baylis, E. *Error Correcting Codes*, Chapman & Hall: London, 1998; pp. 38-42.

5. Schroeder, M. *Fractals Chaos, Power Laws*, W. H. Freeman: New York, 1991; pp. 334-340.

6. Štambuk, N. Fractal model of the hemato-ocular barrier:Verhulst dynamics and mathematical modelling of the inflow and outflow values. *Mathl. Comput. Modelling*, **1990**, 14, 565-570.

7. Steinhaus, H. Nowa wlasnosi mnogosci Cantora. *Wektor*, **1917**, 1-3. (A new property of the Cantor set. *Hugo Steinhaus,Selected Papers*. Polish Academy of Sciences, Institute of Mathematics: Warszava, 1985, pp 205-207.

8. Tufillaro, N.B.; Abbott, T.; Reilly, J. *An Experimental Approach to Nonlinear Dynamics and Chaos*. Addison-Wesley: Redwood City, 1992, pp 96-231

9. Gardner, M. Mathematical games. *Sci. American*, **1972**, August, 106-109.

10. Siemion, I.Z. The regularities of the changes of amino acid physico-chemical properties within the genetic code. *Amino Acids*, 1995, 8, 1-13.

11. Holmes, P. Knotted periodic orbits in the suspensions of Smale's horseshoe: Extended families and bifurcation sequences. *Physica D*, **1989**, 40, 42-62.

12. Grafstein, D. Stereochemical origins of the genetic code. *J. Theor. Biol.* 1983, 105, 157-174.

13. Plantz, A. R.; Berman, M. Adoption of the octal number system. *IEEE Trans. Computers*, **1971**, May, 593-597.

14. Gardner, M. *Mathematical Magic Show*, Penguin books: Harmondsworth, 1989; pp. 97-99.

15. Blalock, J.E.; Bost, K.L. Binding of peptides that are specified by complementary RNAs. Biochem. J. **1986**, 234, 679-683.

16. Varshavsky, A. The N-end rule: functions, mysteries, uses. *Proc. Natl. Acad. Sci. USA*, **1996**, 93, 12142-12149.

17. Crick, F.H.C.; Griffith, J.S.; Orgel, L.E. Codes without commas. *Proc. Natl. Acad. Sci. USA*, **1957**, 43, 416-421.

18. Dounce, A.L. Duplicating mechanismfor peptide chain and nucleic acid synthesis. *Enzymologia*, **1952**, 15, 251-258.

19. Rakočević, M.M. The genetic code as a golden mean determined system. *BioSystems,***1998**, 46, 283-291.

20. Peng, C.-K.; Buldyrev, S.V.; Goldberger, A.L.; Havlin, S.; Sciortino, F.; Simons, M.; Stanley, H.E. Long-range correlations in nucleotide sequences. *Nature,* **1994**, 356, 168-170.

21. Ohno, S. Universal rule for coding sequence construction: TA/CG deficiency - TG/CT excess. *Proc. Natl. Acad. Sci. USA,***1988**, 85, 9630-9634.

22. Yomo, T.; Ohno, S. Concordant evolution of coding and noncoding regions of DNA made possible by the universal rule of TA/CG deficiency - TG/CT excess. *Proc. Natl. Acad. Sci. USA,* **1989**, 86, 8452-8456.

23. Mantegna, R.N.; Buldyrev, S.V.; Goldberger, A.L.; Havlin, S.; Peng, C.-K.; Simons, M.; Stanley, H.E. Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Physical Review E,* **1995**, 52, 2939-2950.

24. Czirok, A.; Mantegna, R.N.; Havlin, S.; Stanley, H.E. Correlations in binary sequences and a generalized Zipf analysis. *Physical Review E,* **1995**, 52, 446-452.

25. Wall, R.W. Terminating decimals in the Cantor ternary set. *Fibonacci Quart.,* **1990**, 28, 98-101.

Figure 1. Binary notation of the 4 nucleotide bases based on purine-pyrimidine and strong-weak H bonding principles. Complementary codon pairs and Siemion's one step amino acid mutation ring[10] are defined by means of the hypercube node distances (Table 1) and related permutations of Table 3 with 4 nucleotide families, as shown in Table 2. Dotted line = 1st base permutation, solid line = 2nd base permutation, • = start; 3rd base permutation involves CU and AG pairs.
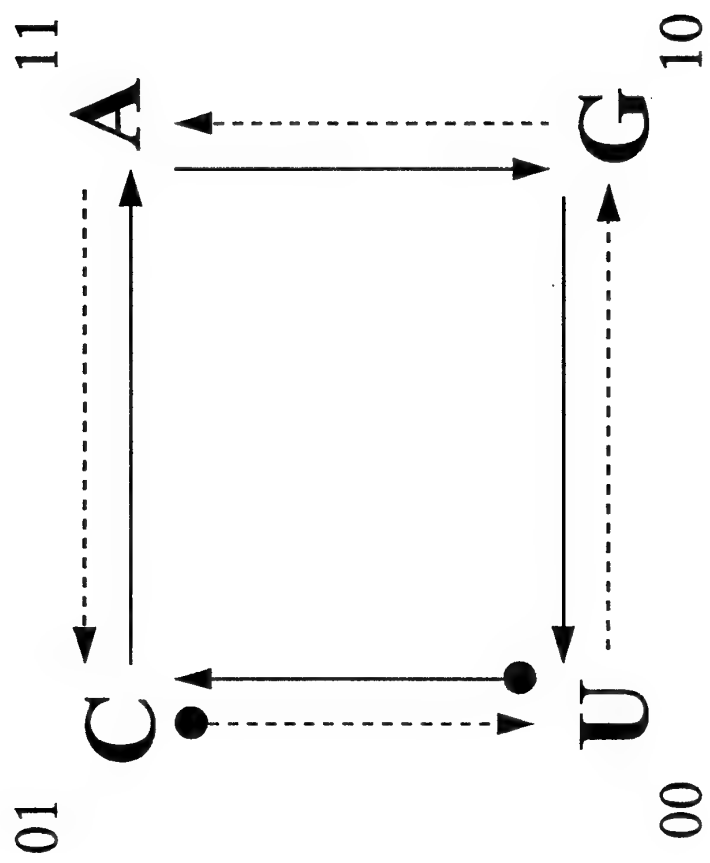
Table 1. Binary and symbolic notation of RNA, DNA and amino acids.

aa ↓	codon ↓	Cantor points	binary notation	*symbolic notation*			aa ↑	codon ↑	Cantor points	binary notation	*symbolic notation*		
F	UUU	0	00 00 00	F	UUU	00 00 00	K	AAA	1	11 11 11	V	GUU	10 00 00
F	UUC	1/243	00 00 01	F	UUC	00 00 01	K	AAG	242/243	11 11 10	V	GUC	10 00 01
L	UUG	2/243	00 00 10	L	UUA	00 00 11	N	AAC	241/243	11 11 01	V	GUA	10 00 11
L	UUA	3/243	00 00 11	L	UUG	00 00 10	N	AAU	240/243	11 11 00	V	GUG	10 00 10
S	UCU	6/243*	00 01 00	S	UCG	00 01 10	R	AGA	237/243*	11 10 11	A	GCG	10 01 10
S	UCC	7/243	00 01 01	S	UCA	00 01 11	R	AGG	236/243	11 10 10	A	GCA	10 01 11
S	UCG	8/243	00 01 10	S	UCC	00 01 01	S	AGC	235/243	11 10 01	A	GCC	10 01 01
S	UCA	9/243	00 01 11	S	UCU	00 01 00	S	AGU	234/243	11 10 00	A	GCU	10 01 00
C	UGU	18/243*	00 10 00	Y	UAU	00 11 00	T	ACA	225/243*	11 01 11	D	GAU	10 11 00
C	UGC	19/243	00 10 01	Y	UAC	00 11 01	T	ACG	224/243	11 01 10	D	GAC	10 11 01
W	UGG	20/243	00 10 10	ochre	UAA	00 11 11	T	ACC	223/243	11 01 01	E	GAA	10 11 11
opal	UGA	21/243	00 10 11	amber	UAG	00 11 10	T	ACU	222/243	11 01 00	E	GAG	10 11 10
Y	UAU	24/243*	00 11 00	W	UGG	00 10 10	I	AUA	219/243*	11 00 11	G	GGG	10 10 10
Y	UAC	25/243	00 11 01	opal	UGA	00 10 11	M	AUG	218/243	11 00 10	G	GGA	10 10 11
amber	UAG	26/243	00 11 10	C	UGC	00 10 01	I	AUC	217/243	11 00 01	G	GGC	10 10 01
ochre	UAA	27/243	00 11 11	C	UGU	00 10 00	I	AUU	216/243	11 00 00	G	GGU	10 10 00
L	CUU	54/243	01 00 00	R	CGU	01 10 00	E	GAA	189/243	10 11 11	S	AGU	11 10 00
L	CUC	55/243*	01 00 01	R	CGC	01 10 01	E	GAG	188/243*	10 11 10	S	AGC	11 10 01
L	CUG	56/243	01 00 10	R	CGA	01 10 11	D	GAC	187/243	10 11 01	R	AGA	11 10 11
L	CUA	57/243	01 00 11	R	CGG	01 10 10	D	GAU	186/243	10 11 00	R	AGG	11 10 10
P	CCU	60/243	01 01 00	Q	CAG	01 11 10	G	GGA	183/243	10 10 11	K	AAG	11 11 10
P	CCC	61/243*	01 01 01	Q	CAA	01 11 11	G	GGG	182/243*	10 10 10	K	AAA	11 11 11
P	CCG	62/243	01 01 10	H	CAC	01 11 01	G	GGC	181/243	10 10 01	N	AAC	11 11 01
P	CCA	63/243	01 01 11	H	CAU	01 11 00	G	GGU	180/243	10 10 00	N	AAU	11 11 00
R	CGU	72/243	01 10 00	P	CCU	01 01 00	A	GCA	171/243	10 01 11	T	ACU	11 01 00
R	CGC	73/243*	01 10 01	P	CCC	01 01 01	A	GCG	170/243*	10 01 10	T	ACC	11 01 01
R	CGG	74/243	01 10 10	P	CCA	01 01 11	A	GCC	169/243	10 01 01	T	ACA	11 01 11
R	CGA	75/243	01 10 11	P	CCG	01 01 10	A	GCU	168/243	10 01 00	T	ACG	11 01 10
H	CAU	78/243	01 11 00	L	CUG	01 00 10	V	GUA	165/243	10 00 11	I	AUG	11 00 10
H	CAC	79/243*	01 11 01	L	CUA	01 00 11	V	GUG	164/243*	10 00 10	M	AUA	11 00 11
Q	CAG	80/243	01 11 10	L	CUC	01 00 01	V	GUC	163/243	10 00 01	I	AUC	11 00 01
Q	CAA	81/243	01 11 11	L	CUU	01 00 00	V	GUU	162/243	10 00 00	I	AUU	11 00 00

*Corresponding positions of the Wall's terminating decimals in the Cantor set.[24] Values denote positions 1/40, 3/40, 1/10, 9/40, 1/4, 3/10, 13/40, 27/40, 7/10, 3/4, 31/40, 9/10, 37/40 and 39/40 respectively.
aa = amino acids; U = T.

Table 2. Rules for the permutation of classic genetic code (Table 3) by the 4 nucleotide families of the first two bases. The algorithm defines: (a) positions of the closest amino acids and (b) places of all amino acid and stop codons in Siemion's one-step mutation ring.[10] Figure 1 presents more details on the binary notation and paths of the circular algorithm.

a.

Base ↓1st/2nd →	00 U	01 C	11 A	10 G
00 U	C	U	G	A
01 C	U	G	A	C
11 A	G	A	C	U
10 G	A	C	U	G

b.

2nd →	U	2 →	C	2 →	A	2 →	G
1st	3rd	1↓	3 →	1↓	3 →	1↓	3 →
↓ C	→ G, A, U, C	U	C, U, A, G	G	G, A, U, C	A	C, U, A, G
↓ U	→ C, U, A, G	G	G, A, U, C	A	C, U, A, G	C	G, A, U, C
↓ G	→ G, A, U, C	A	C, U, A, G	C	G, A, U, C	U	C, U, A, G
↓ A	→ C, U, A, G	C	G, A, U, C	U	C, U, A, G	G	G, A, U, C

Table 3. Classic table of the genetic code.

Base 1st/2nd	U	C	A	G	3rd Base
U	F	S	Y	C	U
	F	S	Y	C	C
	L	S	ochre*	opal*	A
	L	S	amber*	W	G
C	L	P	H	R	U
	L	P	H	R	C
	L	P	Q	R	A
	L	P	Q	R	G
A	I	T	N	S	U
	I	T	N	S	C
	I	T	K	R	A
	M**	T	K	R	G
G	V	A	D	G	U
	V	A	D	G	C
	V	A	E	G	A
	V	A	E	G	G

*stop codons
**start

Table 4. Siemion's one-step mutation ring of the genetic code. Horizontal bars separate three periods of the code to A, C and U family. Italics denote G family codons which are distributed in different periods (third base: Y = pyrimidine, R = purine).

Amino acid ↓	Codon ↓
GGR	*G*
AG period	
GAR	E
GAY	D
AAY	N
AAR	K
CAR	Q
CAY	H
UAY	Y
UAR	stop
UGG	*W*
UGA	*stop*
UGY	*C*
CG period	
UCY	S
UCR	S
GCR	A
GCY	A
ACY	T
ACR	T
CCR	P
CCY	P
CGY	*R*
CGR	*R*
UG period	
CUR	L
CUY	L
UUY	F
UUR	L
GUR	V
GUY	V
AUY	I
AUA	I
AUG	M
AGR	*R*
AGY	*S*
↓	↓
GGY	*G*

Table 5. Horseshoe map representations of 2 binary triplets (or 2 octal numbers, Table 6) extract classic genetic code pattern (Table 3) and define codon mappping based on the unit square (Figure 1) transformation.

Base position	1st	1st	1st	1st	1st	1st	1st	1st	Base position	
U C .100	S	A	T	P	H	N	D	Y	A	U
C C .101	S	A	T	P	H	N	D	Y	A	C
A C .111	S	A	T	P	Q	K	E	ochre*	A	A
G C .110	S	A	T	P	Q	K	E	amber*	A	G
G U .010	L	V	M**	L	R	R	G	W	G	G
A U .011	L	V	I	L	R	R	G	opal*	G	A
C U .001	F	V	I	L	R	S	G	C	G	C
U U .000	F	V	I	L	R	S	G	C	G	U
3rd / 2nd	000.	100.	110.	010.	011.	111.	101.	001.	2nd / 3rd	
↑↓ ↑	U→	G→	A→	C→	←C	←A	←G	←U	↓ ↑↓	

*stop codons
**start

Table 6. Binary coding of the genetic code and protein structure may be transformed into octal coding system based on 3-dimensional cube

permutations. The octal coding system has several advantages, e.g. database compression or simple two–dimensional map projection.

Codon	aa	code	Codon	aa	code	Codon	aa	code	Codon	aa	code	Codon	aa	code	Codon	aa	code	Codon	aa	code	Codon	aa	code
UUU	F	00	UGC	C	11	CUG	L	22	CGA	R	33	GCU	A	44	GAC	D	55	ACG	T	66	AAA	K	77
UUU	F	00	UGC	C	11	CUG	L	22	CGA	R	33	GCU	A	44	GAC	D	55	ACG	T	66	AAA	K	77
UUC	F	01	UGG	W	12	CUA	L	23	CAU	H	34	GCC	A	45	GAG	E*	56	ACA	T*	67			
UGU	C*	10	CUC	L*	21	CGG	R	32	GUA	V	43	GAU	D	54	ACC	T	65	AAG	K	76			
UUG	L	02	UGA	opal	13	CCU	P	24	CAC	H*	35	GCG	A*	46	GAA	E	57						
CUU	L	20	CGC	R*	31	GUG	V*	42	GGA	G	53	ACU	T	64	AAC	N	75						
UUA	L	03	UAU	Y*	14	CCC	P*	25	CAG	Q	36	GCA	A	47									
CGU	R	30	GUC	V	41	GGG	G*	52	AUA	I*	63	AAU	N	74									
UCU	S*	04	UAC	Y	15	CCG	P	26	CAA	Q*	37												
GUU	V	40	GGC	G	51	AUG	M	62	AGA	R	73												
UCC	S	05	UAG	amber	16	CCA	P	27																
GGU	G	50	AUC	I	61	AGG	R	72																
UCG	S	06	UAA	ochre	17																		
AUU	I	60	AGC	S	71																		
UCA	S	07																					
AGU	S	70																					

*Terminating Cantor set decimals[24]

aa = amino acids

Table 7. Circular coding patterns of the 4 colour necklace algorithm. The necklace consists of three horizontal frames with vertically hanging triplets (made of 1, 2 or 3 colours). It is shown that permutations of triplets in horizontal and in vertical direction make 8 collections of 3 colours 12 collections of 2 colours and 4 triplets of 1 colour. The exact location of each triplet in frames given in Table 8 is indicated. In all tables U denotes U or T situation, i.e. the model is valid for both RNA and DNA coding. Wst = opal, Yst = amber or ochre stop codons.

a)

aa	Codon	Frame	aa	Codon	Frame
G	GGU	1	P	CCA	2
V	GUG	2	H	CAC	3
W	UGG	3	T	ACC	1
			complements ⇔		
N	AAC	1	L	UUG	2
T	ACA	2	C	UGU	3
Q	CAA	3	V	GUU	1

b)

aa	Codon	Frame	aa	Codon	Frame
G	GGC*	1	P	CCG*	2
A	GCG	2	R	CGC	3
R	CGG*	3	A	GCC*	1
			complements ⇔		
N	AAU*	1	L	UUA*	2
I	AUA	2	Y	UAU	3
Yst	UAA*	3	I	AUU*	1

c)

aa	Codon	Frame	aa	Codon	Frame
F	UUC*	1	K	AAG*	2
S	UCU	2	R	AGA	3
L	CUU*§	3	E	GAA*§	1
			complements ⇔		
P	CCU*	2	G	GGA*	3
L	CUC§	3	E	GAG§	1
S	UCC*	1	R	AGG*	2

d)

aa	Codon	Frame	aa	Codon	Frame
Y	UAC	1	M	AUG	2
T	ACU*	2	Wst	UGA*	3
L	CUA*	3	D	GAU*	1
			complements ⇔		
R	CGU*	3	A	GCA*	1
V	GUC*	1	Q	CAG*	2
S	UCG	2	S	AGC	3

e)

aa	Codon	Frame	aa	Codon	Frame
I	AUC§	1	Yst	UAG§	2
S	UCA*	2	S	AGU*	3
H	CAU*	3	V	GUA*	1

<div align="center">complements ⟺</div>

aa	Codon	Frame	aa	Codon	Frame
A	GCU	1	R	CGA	3
L	CUG	2	D	GAC	1
C	UGC*	3	T	ACG*	2

f)

aa	Codon	Frame	aa	Codon	Frame
F	UUU	1	K	AAA	1
P	CCC	2	G	GGG	3

<div align="center">complements ⟺</div>

*amino acid (aa) - amino acid dimer assigned codon inversion (exo–endo, exo–endo)[3,12]

§pairs with stop codons and exo–exo, endo–endo inversion[3,12]

Table 8. Necklace model of the genetic code. In horizontal directions we observe circular coding patterns of 3 necklace frames that make Crick'

comma-less code while vertical directions define 64 hanging codons of the evolutional code (a-d). *Wst* = opal, *Yst* = amber or ochre stop codons.

Arrangement of codons in frames according to their projection on the Cantor set transforms each frame in such way that when one letter shift is

performed the next frame is automatically retrieved (a-d).

a)

Frame																					→
m₁ →A GCU	F UUC	Y UAC	L CUC	L CUG	Q CAG	V GUU	V GUC	V GUA	A GCC	G GGU	G GGC	D GAU	D GAC	E GAG	E GAA	I AUU	I AUC	T ACC	N AAU	N AAC	→ m₁ A ↓m₂
m₂ →T CU	S UCU	T ACU	S UCC	C UGC	S AGC	L UUG	S UCG	Yst UAG	P CCG	V GUG	A GCG	M AUG	T ACG	R AGG	K AAG	L UUA	S UCA	P CCA	I AUA	T ACA	→ m₂ AG ↓m₃
m₃ →S U	L CUU	L CUA	P CCU	A GCU	A GCA	C UGU	R CGU	S AGU	R CGC	W UGG	R CGG	Wst UGA	R CGA	G GGA	R AGA	Y UAU	H CAU	H CAC	Yst UAA	Q CAA	S→m₃ AGC↓m₁

b)

Frame																					→
m₁→L ⇐1 CUU	S UCU	T* ACC	S UCC	C UGC	R* AGG	L UUG	S UCG	Yst UAG	P CCG	V GUG	A GCG	M AUG	T ACG	R AGG	K* AAA	L UUA	S UCA	P CCA	I AUA	T̂ ACA	G ↓m₂
m₂→A†,T CU	S UCU	T ACU	S UCC	C UGC	S AGC	L UUG	S UCG	Yst UAG	P CCG	V GUG	A GCG	M AUG	T ACG	R AGG	K AAG	L UUA	S UCA	P CCA	I AUA	T ACA	→ m₂ AG↓m₃

c)

Frame																					→
m₂→V,I† ⇐1 UU	L* CUA	L* CUU	P CCU	A* GCA	A* GCU	C UGU	R CGU	S* AGC	R* CGA	W UGG	R* CGG	Wst UGA	R CGA	G GGA	S* AGU	Y UAU	H* CAC	Q* CAA	Yst UAA	H* CAA	GC ↓m₃
m₃→A,S† U	L CUU	L CUA	P CCU	A GCU	A GCA	C UGU	R CGU	S AGU	R CGC	W UGG	R CGG	Wst UGA	R CGA	G GGA	R AGA	Y UAU	H CAU	H CAC	Yst UAA	Q CAA	S→m₃ AGC↓m₁

d)

Frame																					→
m₃→A,S† ⇐1 C	F UUC	Y UAC	L* CUG	L* CUC	H* CAU	V* GUC	V* GUC	V* GUA	A* GCU	G* GGU	G* GGC	D* GAC	D* GAU	E* GAG	E* GAA	I* AUC	I* AUU	T* ACU	N* AAC	K* AAA	A GCU↓m₁
m₁→A GCU	F UUC	Y UAC	L CUC	L CUG	Q CAG	V GUU	V GUC	V GUA	A GCC	G GGU	G GGC	D GAU	D GAC	E GAG	E GAA	I AUU	I AUC	T ACC	N AAU	N AAC	→ m₁ A ↓m₂

*letter changes that occur during the transformations (see Results and Discussion)

IMMUNOHISTOCHEMICAL ANALYSIS AND PROGNOSTIC VALUE OF CATHEPSIN D DETERMINATION IN LARYNGEAL SQUAMOUS CELL CARCINOMA

Sven Seiwerth [a], Nikola Štambuk [b], Paško Konjevoda [c], Nikola Mašić [b], Vasilj Ankica [a], Miljenko Bura [d], Ivo Klapan [d], Spomenka Manojlović [a], Davor Đanić [e]

[a] Institute of Pathology, Medical Faculty University of Zagreb, Croatia,

[b] Rudjer Bošković Institute, Zagreb, Croatia,

[c] Department of Pharmacology, Medical Faculty University of Zagreb, Croatia

[d] ENT Clinic, Medical Faculty University of Zagreb, Croatia

[e] ENT Department County Hospital, Slavonski Brod, Croatia

Correspondence to: Nikola Štambuk, Rudjer Bošković Institute, Bijenička cesta 54, HR-10001 Zagreb, Croatia. Phone: +385 1 4680 193; Fax: +385 1 4680 094; e-mail: stambuk@rudjer.irb.hr

SUMMARY

Cathepsin D, a protease with the capability of degrading matrix proteins, is implicated in the process of breast and colorectal cancer invasion and metastasis. Biochemical studies in laryngeal cancer have shown a potential prognostic significance of cathepsin D content determination. We studied immunohistochemical positivity of cathepsin D in tumor epithelium and stroma of 61 surgical specimens of squamous cell laryngeal cancer. Immunohistochemical reaction was quantitatively assessed using a PC-based image analysis system SFORM-VAMS. The results were correlated to clinical and morphological parameters and survival. Immunohistochemical positivity was noted in neoplastic cells and tumor stroma. Significant prognostic value for cathepsin D was established separately for epithelial tumor component and tumor stroma using log-rank test, the Cox proportional hazards regression model and C4.5 machine learning system. In all groups, patients above the median cathepsin D staining showed significantly shorter survival time. C4.5 machine learning system extracted cut-off values for the decision tree that defines the probabilities of patients survival and death with high sensitivity (92.8% alive, 73.6% dead), 100% specificity and 86.9% accuracy. This makes immunohistochemical cathepsin D estimation an independent prognostic parameter in laryngeal carcinomas within a 5-year period from the time of tumor surgery.

INTRODUCTION

Cathepsin D is a lysosomal acidic protease[1] thought to be closely associated with tumor invasion or metastasis due to its capability of degrading extracellular matrix[2]. In histopathological and clinical studies, overexpression of cathepsin D was connected with aggressive tumor behaviour in different neoplastic diseases[3,4,5]. Immunochemically the presence of cathepsin D was demonstrated in normal laryngeal mucosa and in primary laryngeal squamous cell carcinomas (SCC)[6]. Recently, a study using radioimmunoasay correlated high cathepsin D content with a poor prognosis, independent of lymph node status[7]. Immunohistochemically cathepsin D was demonstrated in neoplastic and normal laryngeal mucosal cells as well as in stromal macrophages[8]. Aim of our study was to correlate immunohistochemical expression of Cathepsin D to clinical and morphological parameters as well as survival in a laryngeal SCC. We compared the analysis of the results by means of standard log-rank test and Cox proportional hazards regression model to the evaluation of the results by means of C4.5 machine learning system that extracts the decision tree for the classification of the patients survival.

MATERIAL AND METHODS

Patients

We investigated 61 consecutive cases of previously untreated laryngeal SCC patients (males and smokers), without detectable distant metastases, randomly selected from our archives. Figure 1 and 2 illustrate data concerning TNM stages [9] and histopathologic type [9].

Immunohistochemistry

The tissue samples were collected from laryngectomy specimens, fixed in 10% buffered formalin, routinely processed and embedded in paraffin. 3 μm sections were mounted on silanised slides and stained immunohistochemically with an anti-cathepsin D antibody (DAKO, Glostrup) using avidin-biotin method (ABC, Vector, Burlingham) according to manufacturer's specification (Picture 1, cathepsin D = dark brown staining). All the slides were stained in one batch, by one technical assistant.

Image analysis

Immunoreactivity was analysed by two of the authors (SS and AV) on a Leitz Diaplan microscope, using a PC based image analysis system SFORM-VAMS (Zagreb, Croatia; http://www.vams.com)[10] and a CCD camera (JVC TK 1270). Background lightening was kept constant and uniform, and a standard blue filter was used. The immunoreactive area was accessed as percentage of total area analysed (4 fields, objective ×25). Immunoreactivity was analysed separately for epithelial and stromal cells.

Data analysis

Analysed variables were: cathepsin D immunoreactivity (separately for tumor stroma and tumor epithelium), histopathologic grade (grade I and II together), clinical TNM stage and survival. Analysis was made using Kaplan-Meier curves and log-rank test[11], the Cox proportional hazards regression model[12] and C4.5 decision tree learning algorithm[13-18].

C4.5 machine learning program

The program C4.5 is a successor of the basic ID3 deccision tree learning algorithm[13,15,18]. C4.5 generates the classifier in the form of decision tree with elements being either leafs or decision

nodes[13-18]. The leaf shows a class and the decision node specifies the test to be implemented on an attribute value, with one branch and subtree for each possible result of the test[13,15-18]. The starting node is the root node and a tree is used to predict a case by starting at the root and moving through the tree until the leaf is encountered[13,15,17,18]. The C4.5 algorithm presumes existence of appropriate number of learning examples described by set of attributes and by classes representing conditions[13,15,18]. It searches for the most informative attribute according to the gain criterion and constructs decision tree[13,15-18]. This search is based on Shannon's measure of information[13,15,17]. Pruning is used to reduce the decision tree, i.e. for producing more comprehensible structure without compromising accuracy on unseen cases [13, 15,17,18]. For any tree, all paths lead to a leaf corresponding to a decision rule that is a logical conjunction of various tests[15, 18]. If there are multiple paths for a given class, then the paths represent logical disjunctions[13,15,17,18]. All paths are mutually exclusive [13,15,18]. For any new case, one and only one path in the tree will always have to be satisfied[13,15,18].

Sensitivity, specificity and accuracy of the procedures were obtained in a standard way[19]. Predictivity, i.e. reliability of the classifier predictions was calculated as a ratio of the number of true predictions to the size of appropriate prediction class (*alive* or *dead*)[13,15,17].

RESULTS

Histopathological analysis

Follow up of the patients was from 4 to 108 months with a median of 60 months. 42 patients were censored (group *alive*) and 19 completely observed (group *dead*)[12]. Cathepsin D immunoreactivity was histologically observed in normal and neoplastic tissue. In normal laryngeal mucosa next to tumor, diffuse, weak cytoplasmatic positivity was noted. Scarce

reactivity was also present in stromal macrophages. Neoplastic epithelial cells showed mostly diffuse positivity ranging from occasional to majority of cells, with a slightly stronger expression in more dedifferentiated cells. In the tumor stroma abundant immunoreactive cells (macrophages) were noted. Scarce apical reaction in the stroma of the salivary gland cells was also noted.

Kaplan-Meier curves and log-rank test

In the tumor epithelium cathepsin reactivity ranged from 0.06% to 5.63% with a median of 1.35% while in the stroma its range was from 0.62% to 42.02% with a median of 9.86% (Figure 3, 4.). We compared cathepsin D expression with conventional prognostic factors. There was no significant correlation of cathepsin D immunoreactivity with clinical TNM stages (Spearman Rank Order Correlation = 0.24 and $p < 0.05$) as well as with histopathological grading (Spearman Rank Order correlation = 0.08 and and $p < 0.05$). Immunoreactivity for cathepsin D, in both tumor epithelial cells and stroma respectively, showed strong influence on patient survival (Figures 5, 6). Clinical TNM status (Chi-square = 3.896, df = 3, $p = 0.273$) and a histological grade (Chi-square = 3.739, df = 2, $p = 0.154$) showed no significant influence on patient survival.

The Cox proportional hazards regression model

The Cox proportional hazards regression model suggested that only cathepsin D immunoreactivity in epithelial cells have statistically significant effect ($p < 0.05$, Table 1).

C4.5 machine learning system

For the analysis with the C4.5 classifier the patients were divided in two groups (*dead* or *alive* after 60 months, i.e. 5 years following the surgical procedure). A decision tree extracted

cathepsin D epithelial and stromal tumor cell staining as the most significant for the classification of the *alive* or *dead* groups of patients, i.e. the prognosis. The classification rules depicted in the Figure 7. can be read as follows:

1. The first rule says that if the value of RED (epithelial cathepsin D) is less then or equal to 2.33% then a patient belongs to a group *alive*.

2. If that rule is not satisfied (i.e. RED - epithelial cathepsin D is more than 2.33%) then the group is alive when BLUE - stromal cathepsin D is above 38%.

3. The group is *dead* when RED - epithelial cathepsin D is more than 2.33% and BLUE - stromal cathepsin D equal or less then 38%.

The test showed high sensitivity by accurately predicting 5-year survival following the surgical procedure in 92.8% of the patients with laryngeal squamous cell carcinoma (Figure 7). The classifier's sensitivity in predicting the death due to the tumor progression, within 5-year period following the surgery, was satisfactory 73.6% (Figure 7). The specificity of the test was 100% since the decision tree evaluation is made on the tumor tissue, absent in the normal laryngeal immunohistochemical sample. Therefore all patients without tumor have *a priori* negative test result with respect to specificity evaluation[19]. The accuracy of the test was also high (86.9%) and the reliability of classifier's prediction (predictivity) was 88.6% for the group *alive* and 82.3% for the group *dead* (Figure 7).

DISCUSSION

Expression of cathepsin D was analysed in different malignancies such as breast[3,20,21], melanoma [22], colorectal[23], head and neck cancer[24] as well as in childhood or nervous system neoplasm [25,26]. It was suggested that cathepsin D can play a role in tumor cell proliferation by

growth factor activation or promote tumor invasion and metastasis by activating proteolytic enzymes [1,2].

In laryngeal carcinomas different studies established by immunometric assays a higher cathepsin D content in tumor tissue samples, as compared with normal laryngeal mucosa[6,7]. Immunohistochemically strong reactivity was demonstrated for cathepsin D in tumor cells and in tumor stroma macrophages that infiltrate the tissue. In this context very high stromal cathepsin D value linked to the patients survival in subgroup of patients (Figure 7), may be due to the enhanced local immune response to tumor antigens[27].

C4.5 decision tree learning algorithm (Figure 7) was superior to Cox's model (Table 1), regarding the analysis of data structure, since it extracted cut-off values of both epithelial and stromal cathepsin D content relevant for the survival. Similarly to C4.5, log-rank test showed the statistical significance of epithelial and stromal cathepsin D content for patients survival (Figures 5, 6), but failed to explain good prognosis for the subgroup of the patients with extremely high stromal cathepsin D content.

Stromal cathepsin D values of < 38%, linked to the tumor progression (Figure 7) probably reflect enhanced activity of protease concerning the metastasis[1,2] and less pronounced immune response. It is worth mentioning that, with respect to C4.5 based classification, the cut-off prognostic value of stromal cathepsin D is the descendant node of the best predicting attribite[18] (i.e. epithelial tumor cathepsin D) which represents the most informative and root node of the tree[15,18]. Best attribute defining of tumor epithelial cathepsin D by means of C4.5 classifier is in agreement with the fact that laryngeal sqamous cell carcinoma is an epithelial neoplasm.

Prognostic significance was recently assumed for radiometrically measured cathepsin D levels[6,27,28], however, to our knowledge this is the first immunohistochemical study demonstrating strong prognostic significance of cathepsin D in laryngeal cancer. Besides being highly prognostic, this type of immunochemical test is relatively cheap and easy to perform which makes the combination of quantitative immunohistochemical analysis of cathepsin D and C4.5 based data classification a potent prognostic tool in laryngeal cancer patients.

Although the surgical treatment is *condicio sine qua non* in the therapy of laryngeal squamous cells carcinoma it seems that cathepsin D content in histopathology samples of the tumor cells represents an important predictive factor for tumor recidives and aggressive behaviour. In this study the data analysis was not influenced by other therapeutic procedures (e.g. chemotherapy or radiotherapy) due to the fact that surgery is a primary therapeutic procedure for the type and stage of laryngeal neoplasm we observed (Figure 1). It is worth mentioning that all of our patients were males and smokers.

In our study expression and localisation of cathepsin D immunoreactivity correlated with data obtained by others for laryngeal neoplasms [27,28]. From our results, cathepsin D seems to be an independent prognostic marker in primary laryngeal carcinomas, which confirms the hypothesis of Marsigliante *et al.*[28]. Decision tree extracted by means of C4.5 classifier was shown to be a valuable tool to define highly sensitive, specific, accurate and predictive cut-off values for immunohistochemical cathepsin D data. It remains an open question if this method of analysis could be applied to other tumors (e.g. breast,colorectal and gactric cancer)[3-5,21,29-31] with established link between aggressive neoplastic behaviour and cathepsin D content.

REFERENCES AND NOTES

1. Capony, F.; Morrisset, M.; Barrett, A.J.; Capony, J.P.; Broquet, P.; Vignon, F.; Chambon, M.; Louist, P.; Rochefort, H. Phosphorylation, glycosilation and proteolytic activity of the 52-kD estrogen-induced protein secreted by MCF7 cells. *J.Cell Biol.* **1987**, 104, 253-262.

2. Montcurrier, P.; Mangeant, PH.; Salazar, G.; Morisset, M.; Sahuquet, A.; Rochefort, H. Cathepsin D in breast cancer can digest extracellular matrix in large acid vesicles. *Cancer Res.* **1990**, 50, 6045-6054.

3. Kute, T.E.; Shao, Z.M.; Sugg, K.; Long, R.T.; Russell, G.B.; Case, D. Cathepsin D as prognostic indicator for node-negative breast cancer patients using both immunoassay and enzymatic assay. *Cancer Res.* **1992**, 52, 5198-5203.

4. Rochefort H.; Capony, F.; Garcia, M. Cathepsin D: A protease involved in breast cancer metastasis. *Cancer Metastasis Rev.* **1990**, 9, 321-331.

5. Leto, G.; Gebbia, N.; Rausa, L.; Tuminello, FM. Cathepsin D in the malignant progression of neoplastic diseases. *Anticancer Research,* 1992, 12, 235-240.

6. Marsigliante, S.; Resta, L.; Leo, G.;Mazzotta, D.; d'Amore, R.; Biscozzo, L.; Storelli, C. Expression of cathepsin D in malignant an in the corresponding non-malignant node-negative laryngeal samples: correlation with receptors for androgen, glucocorticoid, oestrogen and progesterone. *Cancer Letters,* **1993**, 68, 135-142.

7. Maurizi, M.; Almadori G.; Cadoni, G.; Scambia, G.; Ottaviani, F.; Ferrandina, G.; Paludetti, G.; D'Abramo, G.; Manusco, S. Cathepsin D concentration in primary laryngeal cancer: correlation with clinico-pathological parameters, EGFR status and prognosis. *Int. J. Cancer,* **1996**, 69, 105-109.

8. Ferrandina, G.;Scambia, G.; Benedetti, P.; Almadori, G.; Paludetti, G.; Cadoni, G.; Distefano, M.; Maurizi, M.; Mancuso, S. Cathepsin D in primary squamous laryngeal tumors: correlation with clinico-pathological parameters and receptor status. *Cancer Letters*, **1992**, 67, 133 -138.

9. American Joint Commite on Cancer. Manual for Staging on Cancer; JB Lippincot, Philadelphia, 1988, p 43.

10. Konjevoda, P.; Nasić, M.; Ćurković, T.; Sikirić, P.; Seiwerth, S.; Štambuk, N. Effects of BPC157 on the healing of corneal lesions. In *Uveitis Today*; Ohno, S.; Aoki, K.; Usui, M.; Uchio. E., Eds.; Elsevier: Amsterdam, 1998; pp 311-314.

11. Bartels, P.H.; Thompson, D.; Weber, J.E. In *Image Analysis - A primer for Pathologist*; Marchevsky, A.M.; Bartels P.H., Eds.; Raven Press: New York, 1994; Chapter 7, p 207.

12. Afifi, A.A.; Clark, V. *Computer-aided Multivariate Analysis*; Chapman & Hall: London, 1996; Chapter 13, p 306.

13. Quinlan, J.R. Induction of Decision Trees. *Machine Learning*, **1986**, 1, 81- 106.

14. Blazadonakis, M.; Moustakis, V.; Charissis, G. Deep assessment of machine learning techniques using patient treatment in acute abdominal pain in children. *Artif. Intell. Med.*, **1996**, 8, 527 - 542.

15. Quinlan, J.R. *C4.5: Programs for Machine Learning*, Morgan Kaufmann: San Mateo, 1993; pp 3-29.

16. Quinlan, J.R. Rule induction with statistical data-a comparison with multiple regression. *J. Operat. Res. Soc.*, **1987**, 38, 347-352.

17. Štambuk, N.; Mašić, N.; Brinar, V.; Trbojević-Čepe, M.; Štambuk, V.; Rabatić, S.; Marotti, T.; Šverko, V.; Marušić-Della Marina, B.; Brzović, Z.; Zurak, N.; Svoboda-Beusan, I.; Mažuran, R.; Martinić, R.; Štambuk, A.; Karaman, K.; Gagro, A.; Rudolf,

M.; Malenica, B.; Pokrić, B. Machine learning approach defines immune response during met-enkephalin immunotherapy in multiple sclerosis and optic neuritis. In *Advances in Systems, Signals, Control and Computers, vol. II*; Bajić, V.B., Ed.; International Association for the Advancement of Methods for System Analysis and Design (IAAMSAD) and Academy of Nonlinear Sciences (ANS): South Africa, 1998, pp 302-305.

18. Mitchell, T.M. *Machine Learning*, McGraw-Hill: Singapore, 1997; pp 55-80.

19. Campbell, M.J.; Machin, D. *Medical Statistics*, J. Wiley & Sons: Chicester, 1995; Chapter 3, pp 34-35, 74.

20. Tandon, A.K.; Clark, G.M.; Chamness, G.C.;Chirgwin, .JM.; McGuire, W.L. Cathepsin D and prognosis in breast cancer. *N. Engl. J. Med.*, **1990**, 322, 297-302.

21. Schultz, D.C.; Bazel, S.; Wright, L.M.; Tucker, S.; Lange, M.K.; Tachovsky, T.; Lomgo, S.; Niedbala, S.; Alhadeff, J.A. Western blotting and enzymatic activity analysis of cathepsin D in breast tissue and sera of patients with breast cancer and benign breast disease and of normal controls. *Cancer Res.*, **1994**, 54, 48-54.

22. Goldmann, F.J.; Biess, B.; Lippold, A.; Suter, L.; Westhoff, U. Prognostic classification of malignant melanomas by combining clinical, histological, and immunohistochemical parameters. *Oncology*, **1999**, 56, 208 - 214.

23. Tuminello, F.M.; Gebbia, N.; Pizzolanti, G.; Russo, A.;Bazan, V.; Leto, G. Cathepsin D content in colorectal cancer. *Oncology*, **1995**, 52, 237-242.

24. Zeillinger, R.; Swoboda, H.; Machacek, E.; Nekahm, D.; Sliutz, G.; Knogler, W.; Speiser, P.; Swoboda, E.; Kubista, E. Expression of cathepsin D in head and neck cancer. *Eur. J. Cancer*, **1992**, 28A, 1413-1415.

25. Parham, D.;Whitaker, J.N.; Costan, W.; Berard, W. Cellular distribution of cathepsin D in childhood tumors. *Arch. Pathol. Lab. Med.*, **1985**, 109, 250-255.

26. Robson, D.K.; Ironside, J.W.; Reid, W.A.; Bogue, P.R. Immunolocalization of cathepsin D in the human central nervous system and central nervous system neoplasms. *Neuropathol. Appl. Neurobiol.*, **1990**, 16, 39-44.

27. Resta, L.; Fiorella, R.; di Nicola, V.; Marzullo, A.; Boticella, M.A.; Maiorano, E. Cathepsin D in laryngeal carcinoma. Preliminary report. *Bolletino-Societa Italiana Biologia Sperimentale*, **1995**, 71, 257-261.

28. Marsigliante, S.; Biscozzo, L.; Resta, L.; Leo, G.; Mottaghi, A.; Maiorano, E.; Collucci, G.; Storelli, C. Immunohistochemical and immunoradiometric evaluations of total cathepsin-D in human larynx. *Eur J. Cancer*, **1994**, 30B, 51-55.

29. Marsigliante, S.; Mottaghi, A.; Muscella, A.; Cirado, V.; Leo, G.; Storelli, C. Effect of tumor size on the association between PS2 and cathepsin D in primary breast cancer. *Int. J. Oncol.*, **1995**, 6, 69-73.

30. Theodoropoulos, G.E.; Panoussopoulos, D.; Lazaris, A.C.; Golematis, B.C. Evaluation of cathepsin D immunostaining in colorectal adenocarcinoma. *J. Surg. Oncol.*, **1997**, 65, 242-248.

31. Allgayer, H.; Babic, R.; Grutzner, K.U.; Beyer, B.C.M.; Tarabichi, A.; Schildberg, F.W., Heiss, M.M. Tumor-associated proteases and inhibitors in gastric cancer - analysis of prognostic impact and individual risk protease patterns. *Clin. Exp. Metastasis*, **1998**, 16, 62-73.
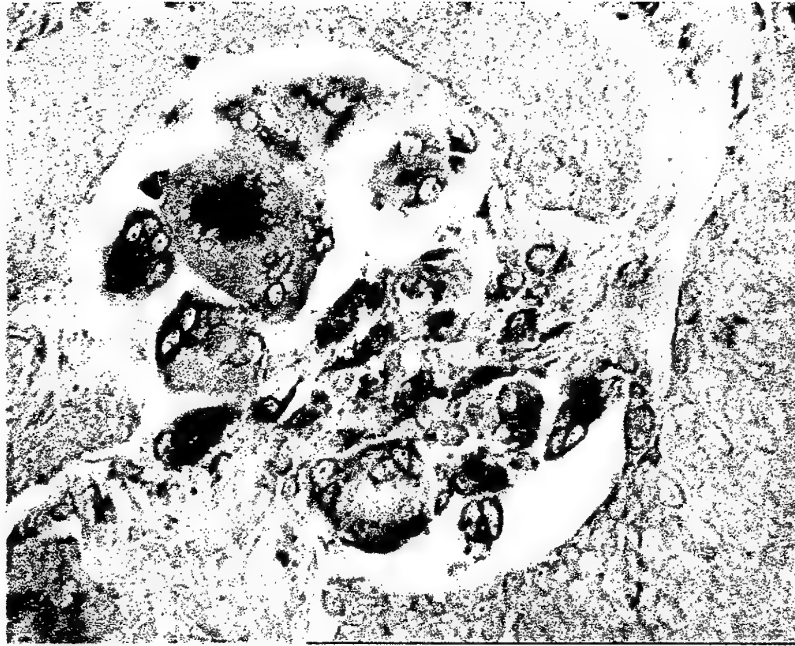
Picture 1. Cathepsin D positivity in tumor cells. (×40)

Fig 1. Distribution of patients by TNM classification
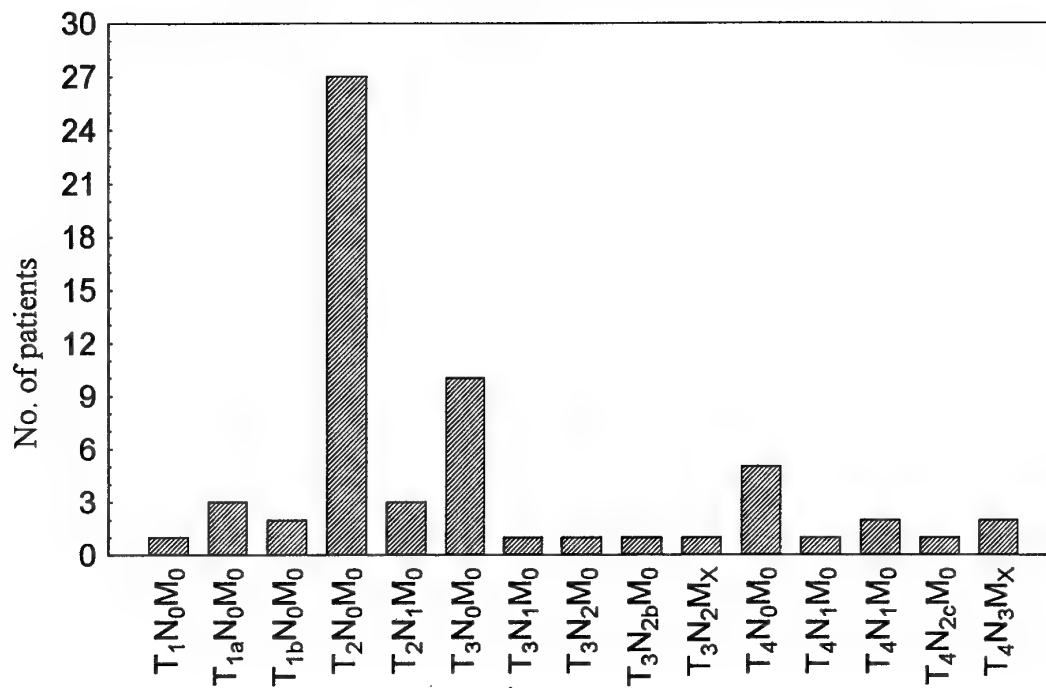
Fig 2. Distribution of patients by Gradus (I - IV)
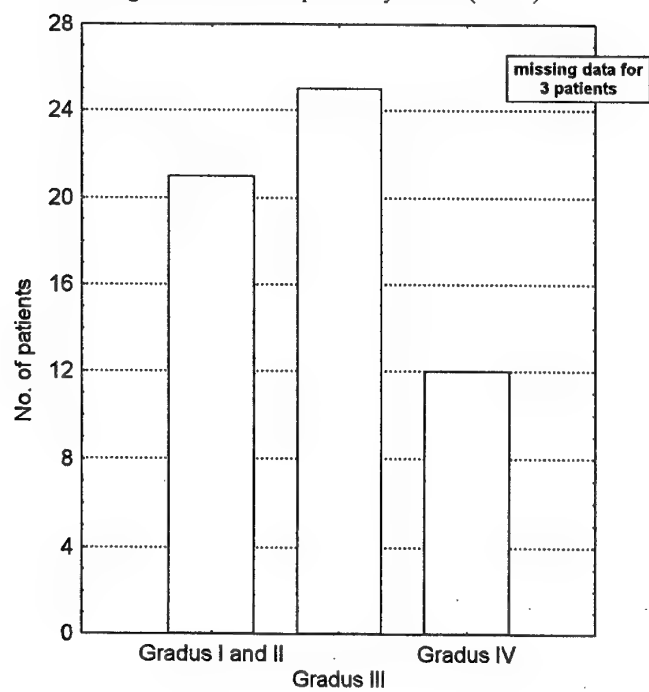
Fig 3. Distribution of patients by cathepsin D positive area (%) in the epithelial component of tumor
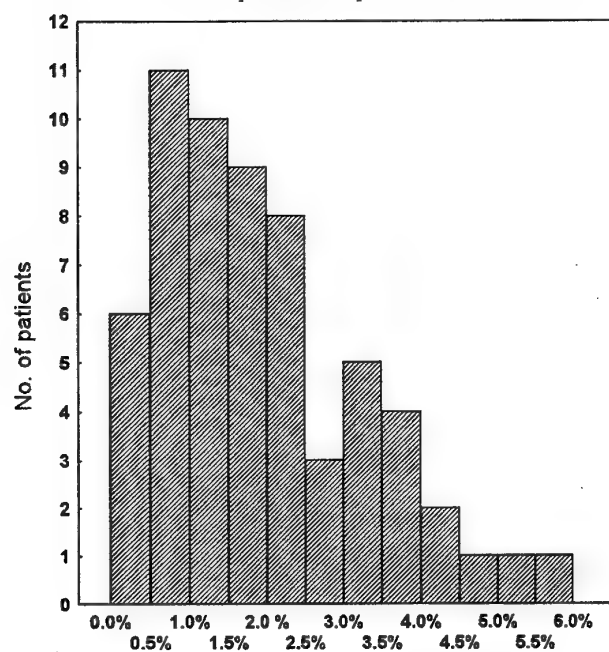
Fig 4. Distribution of patients by cathepsin D
positive area (%) in tumor stroma
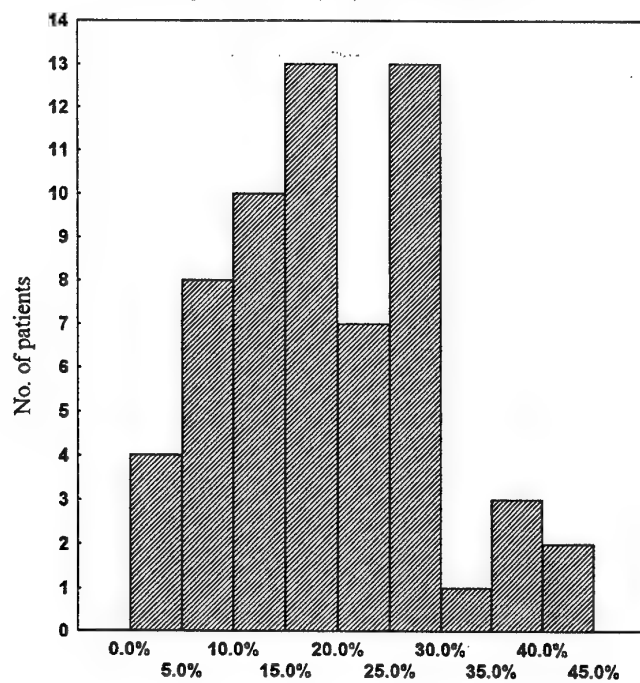
Fig 5. Survival analysis for cathepsin D positive area (%)
in the epithelial component of tumor (Kaplan - Meier)
o Complete response   + Censored observations

Fig 6. Survival analysis for cathepsin D positive area (%)
in tumor stroma (Kaplan - Meier)
○ Complete response   + Censored observations
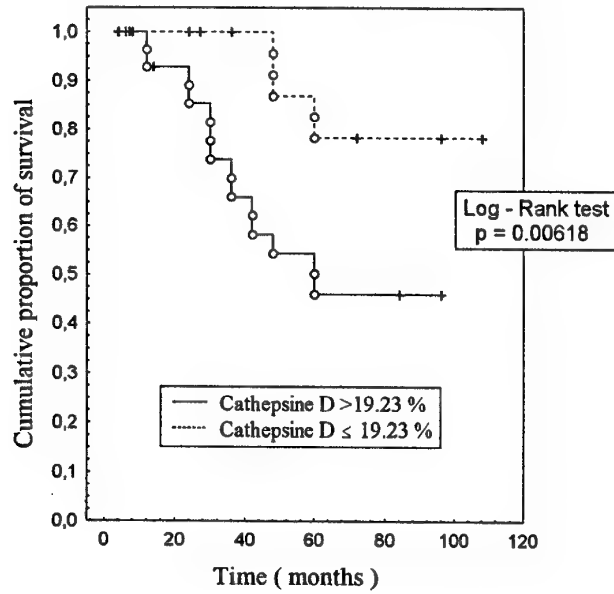
Fig 7. Decision tree obtained by C4.5 machine learning system

SENSITIVITY (alive) = 92.8 % = 39/42
SENSITIVITY (dead) = 73.6 % = 14/19
PREDICTIVITY (alive) = 88.6 % = 39/44
PREDICTIVITY (dead) = 82.3 % = 14/17
SPECIFICITY = 100 %
ACCURACY   = 86.9 %
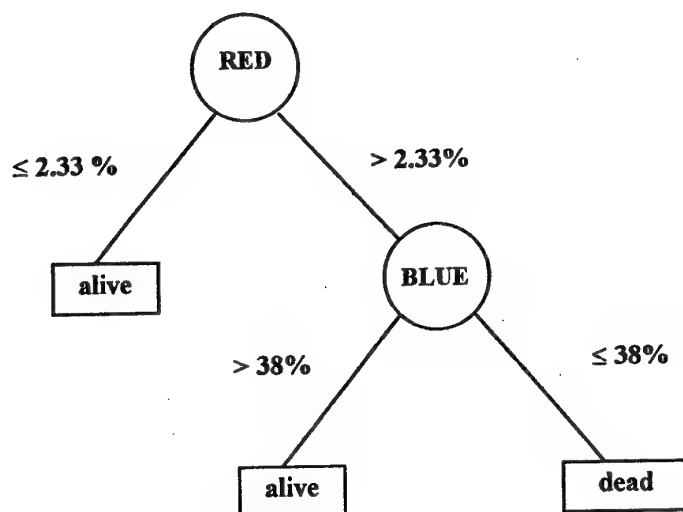
Table 1. Cox proportinal hazard risk model for clinical stage, cathepsin D content in epithelial and stromal component of tumor and histopathological gradus

Cox model	Chi-square = 18.3673	df = 4	p = 0.0015	
Variables	Estimate	Standard error	t-value	two-sided p value
Clinical stage	0.43	0.28	1.51	0.13
Cathepsin D (epithel)	0.83	0.32	2.30	0.009
Cathepsin D (stroma)	-0.04	0.05	0.96	0.46
Histopathological grade	0.16	0.29	0.54	0.59

THE VALUE OF THE URINARY STONE PROMOTERS/INHIBITORS RATIOS IN THE ESTIMATION OF THE RISK OF UROLITHIASIS

Danica Batinić[1], Danko Milošević[1], Nenad Blau[2], Paško Konjevoda[3], Nikola Štambuk[4], Vesna Barbarić[5], Mirna Šubat-Dežulović[6], Ana Votava-Raić[1], Ljiljana Nižić[1] and Kristina Vrljičak[1]

Childrens' University Hospital Šalata, Department of Pharmacology, Clinical Institute for Laboratory Diagnosis, Medical Faculty University of Zagreb, and Rudjer Bošković Institute, Zagreb, and University Childrens' Hospital, Medical Faculty of Rijeka, Croatia, and Childrens University Hospital, Zurich, Switzerland

1. Childrens' University Hospital Šalata

2. Childrens' University Hospital Zurich

3. Department of Pharmacology

4. Rudjer Bošković Institute

5. Clinical Institute for Laboratory Diagnosis

6. Childrens' University Hospital Rijeka

Correspondence to: Danica Batinić, Childrens' University Hospital Šalata, Medical Faculty University of Zagreb, Šalata 4, 10000 Zagreb, Croatia

SUMMARY

An imbalance between urinary promoting and inhibiting factors has been suggested as more important in urinary stone formation than a disturbance of any single substance. To investigate the value of promoter/inhibitor ratios for estimation of the risk of urolithiasis, urinary citrate/calcium, magnesium/calcium oxalate and oxalate/citrate×glycosaminoglycans ratios were determined in 30 children with urolithiasis, 36 children with isolated hematuria, and 15 healthy control children. The cut-off points between normal children and children with urolithiasis, accuracy, specificity and sensitivity for each ratio were determined and compared with those of the 24h-urine calcium and oxalate excretion and urine saturation calculated with the computer program EQUIL2. The neural network application (aiNET Artificial Neural Network, version 1.25) was used for the determination of the cut-off points for the classification of a normal children and urolithiasis group. The best test for differentiating stone formers from non-stone formers proved the aiNET determined cut-off values of oxalate/citrate×glycosaminoglycans ratio. The method showed 97.78 % accuracy, 100% sensitivity and 93.33 % specificity. Two cut-off points between normal and urolithiasis groups were found showing that the children with urolithiasis had the ratio values either above 34.00 or less than 10.16. Increased oxalate excretion was linked to the first cut-off value (34.00) and decreased glycosaminoglycans excretion was typical of the second cut-off value (10.16).

Key words: urinary stones, promoter/inhibitor ratios, risk of urolithiasis, neuronal network, classification

INTRODUCTION

Extensive examination of number of urinary promoting and inhibiting factors was undertaken over years to investigate risk for stone formation. It has been shown that no single promoter or inhibitor can discriminate clearly enough any particular individual as healthy or sick. Combination of factors seems to provide better separation of stone formers from normal subjects. Several ratios between promoting and inhibiting factors, such as calcium $\times$ oxalate/creatinine $\times$ magnesium[1], calcium/citrate[2], magnesium/calcium$\times$oxalate[3], oxalate/calcium[4], citrate/calcium[5], oxalate/citrate$\times$glycosaminglycans[6], as well as more sophisticated methods that take into account the number of urinary compnents[7-14], were used to detect the imbalance between the promoting and inhibiting factors leading to stone formation. We examined 11 single urinary factors potentially promoting or inhibiting crystallization and urine saturation with computer program EQUIL 2 in children with isolated hematuria and overt urolithiasis and compared the findings with the findings of normal healthy children[15]. In our previous report urine saturation was found as the best parameter for the estimation of the relative risk of urolithiasis. However, logistic regression failed correctly to classify 14.59% of the group members[15].

The aim of the present study was to evaluate the value of promoter/inhibitor ratios for estimation of the risk of urolithiasis. Those tests are simpler, easier and cheaper for routine clinical practice than EQUIL2. Citrate/calcium, magnesium/calcium$\times$oxalate and oxalate/citrate$\times$ glycosaminoglycans ratios were chosen for this purpose because all of them take into account the major urinary stone promoting and inhibiting factors. Neuronal networks analysis by means of aiNET Artificial Neural Network (version 1.25) was used to determine the cut-off points between normal and urolithiasis groups. Accuracy, specificity and sensitivity were calculated for each of these ratios

and compared with those of the 24-h urine calcium and oxalate excretion and urine saturation.

PATIENTS AND METHODS

Patients. Thirty children with urolithiasis and 36 children with isolated hematuria were investigated. A group of 15 healthy sex and age matched children without any nephrourological disease or pathological condition that might influence urine composition served as controls.

In children with hematuria, glomerular diseases, urinary infection, urological anomalies and coagulopathy were excluded before entering the study. If a checkup of serum and urine electrolytes revealed hypercalciuria, the known causes of hypercalciuria (renal tubular acidosis, hypercalcemic conditions) were excluded.

In children with urolithiasis ultrasonography and/or urography established the diagnosis. Cystinuria and hyperuricosuria were excluded.

The children were enrolled in the study with informed parental consent.

Urine sampling and analysis. From each child, 24-h urine collections performed on two consecutive days and one urine sample collected from 8 to 10 a.m. on the third day were obtained for analysis. The 24-h urine of the first day served for measuring creatinine, calcium, sodium, potassium, oxalate, phosphate, magnesium, citrate and sulphate. It was collected in a wide-mouthed plastic bottle containing 10 ml 6N hydrochloric acid as preservative. The 24-h urine of the second day was collected in the same way but without the addition of hydrochloric acid in the bottle. It served for measuring chloride, urate, GAGs and creatinine. The 2-hour urine collected on the third day served for ammonium and creatinine measuring. In this sample 500 mg di-potassium-oxalate was added

immediately after voiding to prevent ammonium decomposition.

pH of urine was measured with indicator sticks (Boehringer Mannheim, Germany). Oxalate, citrate and sulphate were measured using a Dionex Series 4000i gradient ion chromatography system (Dionex Co, Sunnyvale, CA, USA)[16]. GAGs were measured by carbazole method [17], ammonium by glutamate dehydrogenase (Da Fonseca-Wollheim)[18] and magnesium by atomic absorption spectrophotometry[19]. The following analyses were done on Olympus AU 800 Analyser: creatinine by standard kinetic Jaffé procedure[20], sodium, potassium and chloride by ion selective electrodes, calcium by the cresolphthalein-complexon method[21], phosphate by molybdate method[22] and uric acid by uricase method[23].

From the values of urinary 24-h volume, pH of urine, calcium, sodium, potassium, chloride, magnesium, phosphate, sulphate, ammonium, urate, oxalate, citrate and creatinine (mmol/L), the urinary calcium oxalate saturation was calculated by the computer program EQUIL 2[14,15]. Also, 24-h urinary excretion expressed as a ratio to the creatinine was calculated for each of the measured urinary components.

Data analysis. Data were presented as medians with minimum and maximum values. Cut-off values between normal children and children with urolithiasis were determined using a neuronal network application (aiNET Artificial Neural Network Version 1.24, Celje, Slovenia) [24,25]. Artificial neural network aiNET is based on a self-organising system, called neural network-like system, and it is very similar to the Kohonen's self organisation process[25]. The algorithms used by aiNET do not require any learning phase and the answers about prediction are obtained almost immediately[24, 25]. When the data is chaotic and there is no possible solution aiNET will suggest a data problem[24, 25].

Accuracy, specificity and sensitivity, as well as 95% confidence interval for determined cut-off were calculated.

RESULTS

Table 1. shows median (minimum - maximum) values of the 24-h urinary excretion of calcium and oxalate, urine saturation, citrate/calcium, magnesium/calcium×oxalate and oxalate/citrate× gycosaminoglycans ratios. Cut-off points between normal children and children with urolithiasis were possible to determine using all variables except a magnesium/calcium×oxalate ratio, which data were too dispersed for such discrimination (Table 2). Children with urolithiasis had urine saturation, calcium/creatinine, oxalate/creatinine and citrate/calcium above 4.70, 0.20, 0.48 and 1.38 respectively. For oxalate/citrate×glycosaminoglycans ratio two cut-off points were found. Children with urolithiasis had the ratio values either above 34.80 or less than 10.16. The most accurate method for discrimination normal from sick children was oxalate/citrate×glycosaminoglycans ratio that showed as 100% sensitive and highly specific with only 6.67% false positive results. Than follow citrate/calcium ratio and urine saturation, the former with better sensitivity and the later with the better specificity. All children with the urolithiasis had at least 3 of 5 examined variables in the range of the pathological values and in 19 out of 30 (63.3%) children all variables showed pathological values (Table 3). On contrary all normal children except 1, had no more than 2 variables in the range of the pathological values. In children with hematuria results were dispersed, although the tendency of having higher number of pathological variables was noticed.

DISCUSSION

It seems reasonable to consider urolithiasis as multifactorial disorder with risk of stone

formation dependant upon a disturbance in the balance of promoting and inhibiting factors. In our

previous study we have shown that urine saturation estimates the relative risk of urolithiasis better

than any single urinary constituent[15]. However, the determination of urine saturation may be

inconvenient for routine clinical practice being time consuming and expensive. Therefore, we tried

to find an easier parameter with high sensitivity to discriminate stone formers from healthy children.

Saturation can also be expressed in terms of ratios between urine concentration of 2 or 3 substances

involved in lithogenesis. Among the ratios examined in this study the best proved oxalate/citrate×

glycosaminoglycans. Baggio et al. first suggested this ratio as simple method for detection of the

imbalance between promoting and inhibiting factors and found abnormally high ratio values in

children with idiopathic urolithiasis[6]. The ratio can differentiate more than 80% of stone formers from

non-stone formers. In the present study not only very high, but also very low values of the ratio were

found in children with urolithiasis in comparison with normal children. The ratio values above upper

limit of normal belonged to patients with increased oxalate excretion, while the ratio values under

lower limit of normal reflected relatively decreased glycosaminoglycans excretion. In our previous

study[15] standard statistical methods (two-way analysis of variance and Tukey HSD test with

correction for unequal N) could not detect influence of glycosaminoglycans on differentiation

between normal children and urolithiasis group. The result of neural network classification (Table 2)

based on the artificial intelligence method of data analysis implies that decreased glycosaminoglycans

values may influence the stone formation in a sub-population of stone-formers. The use of two cut-

off points made it possible to increase accuracy of the oxalate/citrate×glycosaminoglycans ratio in differentiation between stone formers and normal children from 80% to 97.78%. Only one healthy child had pathologically high ratio value due to unexpectedly high glycosaminoglycans excretion, the finding that is difficult to interpret.

The citrate/calcium ratio has proved as very good discriminator between stone formers from normal children, too. Although of somewhat lesser accuracy and specificity than oxalate/citrate× glycosaminoglycans ratio (Table 2), a citrate/calcium ratio has advantage of easy performing in clinical practice.

The study showed once more that the disturbance of more than one of substances involved in lithogenesis must be present for stone formation. All children with urolithiasis had at least 3 pathological parameters, while all but 1 normal child has no more than 2 pathological parameters.

The neural network analysis of the laboratory tests related to the important and common medical problem of urolithiasis proved to be of potential clinical value.

LITERATURE

1.   Tiselius, H.G. Relationship between the severity of renal stone disease and urine composition. *Eur.Urol.* **1979**, 15, 5.

2.   Butz, M.; Schulte, P.; Knispel, H. *Neue Gesichtspunkte zur Interpretation harnchemische Befunde beim Harnsteinleiden.* Verh.Ber.34, Tagg Dtsch Ges Urologie, Hamburg 1982. Springer Verlag: Berlin, Heidelberg, New York, Tokyo, 1983, p 331.

3.   Robertson, W.G. Physico-chemical aspects of calcium stone-formation in the urinary tract. In: *Urolithiasis Research*; Fleisch, H.; Robertson, W.G.;Smith, L.H.; Vahlensieck, W., Ed.; Plenum Press: New York, 1976; p 25.

4.   Robertson, W.G.; Rutherford, A. Aspects of the analysis of oxalate in urine – a review. *Scand J Urol Nephrol*, **1981**, (Suppl.) 53:85.

5.   Perrone, H.C.; Toporovski, J.; Shor, N. Urinary inhibitors of crystallization in hypercalciuric children with hematuria and nephrolithiasis. *Pediatr Nephrol*, **1996**, 10, 435-437.

6.   Baggio, B.; Gambaro, G.; Favaro, S.; Borsatti, A.; Pavanello, L.; Siviero, B.; Zacchello, G.; Rizzoni, G.F. Juvenile renal stone disease: a study of urinary promoting and inhibiting factors. *J Urol*, **1983**, 130, 1133-1135.

7.   Marshall, R.W.; Robertson, W.G. Nomograms for the estimation of the saturation of urine with calcium oxalate, calcium pyrophosphate, magnesium amonium phosphate, uric acid, sodium acid urate, ammonium acid urate and cystine. *Clin. Chim. Acta*, **1976**, 72, 253-260.

8.   Robertson, W.G.; Peacock, M; Marshall, R.W.; Marshall, D.H.; Nordin, B.E.C. Saturation-inhibition index as a measure of thwe risk of calcium oxalate stone formation in the urinary

tract. *N.Engl. J. Med.* **1976**, 294, 249-252.

9. Finlayson, B. Calcium stones: some physical and clinical aspects. In: *Calcium Metabolism in Renal Failure and Nephrolithiasis*; David D.S., Ed.; Wiley: New York, 1977; Chapter 10, p 337-382.

10. Robertson, W.G.; Peacock, M.; Heyburn, P.J.; Marshall, D.H.; Clark, P.B. Risk factors in calcium stone disease of the urinary tract. *Br. J. Urol.* **1978**, 50, 449 - 452.

11. Pak, C.Y.C.; Galosy, R.A. Propensity for spontaneous nucleation of calcium oxalate. Quantitative assessment by urinary FPR-APR discriminant score. *Am. J. Med.* **1980**, 69, 681-689.

12. Pak, C.Y.C.; Skurla, C.; Harvey, J. Graphic display of urinary risk factors for renal stone formation. *J. Urol.* **1985**, 134, 867-870.

13. Tiselius, H.G. Measurement of the risk of calcium oxalate crystallization in urine. *Urol. res.* **1985**, 13, 297-300.

14. Werness, P.; Brown, C.M.; Smith, L.H.; Finlayson, B. EQUIL II: a BASIC computer program for the calculation of urinary saturation. *J. Urol.* **1985**, 134, 1242-1244.

15. Milošević, D.; Batinić, D.; Blau, N.; Konjevoda, P.; Štambuk, N.; Votava-Raić, A.; Barbarić, V.; Fumić, K.; Rumenjak, V.; Stavljenić-Rukavina, A.; Nižić, Lj.; Vrljićak, K. Determination of urine saturation with computer programe EQUIL 2 as a method for estimation of the risk of urolithiasis. *J Chem Inf Comput Sci*, **1998**, 38, 646-650.

16. Classen, A.; Hesse, A. Measurement of urinary oxalate: an enzymatic and ion chromatographic method compared. *J. Clin. Chem. Clin. Biochem.* **1987**, 25, 95 -99.

17. Bitter, T.; Muir, H.M. A modified uronic acid carbazole reaction. *Anal Biochem*, **1962**, 4,

330-334.

18. Da Fonseca-Wollheim, F. Z. Bedeutung von Wasserstoffionenkonzentration und ADP-Zusatz bei der Ammoniakbestimmung mit Glutamatdehydrogenase. *Klin. Chem. Klin. Biochem.* **1973**, 11, 421-425.

19. Hansen, J.L.; Frier, E.F. The measurement of serum magnesium by atomic absorption spectrophotometry. *Am. J. Med. Techol.* **1967**, 33, 158-166.

20. Bonses, R.W.; Taussky, H.H. On the colorimetric determination of creatinine by the Jaffe reaction. *J. Biol. Chem.* **1951**, 158, 581-591.

21. Connerty, H.V.; Briggs, A.R. Determination of serum calcium by means of orthocresolphthalein complexone. *Am. J. Clin. Pathol.* **1966**, 45, 290-296.

22. Drewes, P.A. Direct colorimetric determination of phosphorus in serum and urine. *Clin. Chim. Acta,* **1972**, 39, 81-88.

23. Liddle, L.; Seegmiller, J.E.; Laster, L. Enzymatic spectrophotometric method for determination of uric acid. *J. Lab. Clin. Med.* **1959**, 54, 903-913.

24. aiNET Artificial Neural Network Version 1.25; *Users Manual*;Trubarjeva 42, SI-3000 Celje, Slovenia, Europe, **1995** (e-mail: ainet@siol.net).

25. Grabec, I.; Self-Organisation of Neurons Described by the Maximum-Enthropy Principle. *Biol. Cybern.* **1990**, 63, 403-409.

Table 1. Urinary Promoters and Inhibitors of Crystallization [a]

parameters	normal children			hematuria			urolithiasis		
	md	min	max	md	min	max	md	min	max
urine saturation	1.91	0.76	3.48	4.32	0.74	16.29	19.59	0.72	25.20
calcium/creatinine (mmol/mmol)	0.17	0.08	0.33	0.22	0.11	0.79	0.30	0.08	0.56
oxalate/creatinine (mmol/mol)	46.00	19.00	76.00	51.00	8.00	111.00	75.00	20.00	111.00
citrat/calcium ratio (mmol/mmol)	1.81	0.48	3.59	1.33	0.31	7.74	0.71	0.17	1.38
magnesium/calcium × oxalate ratio (mmol)	1.05	0.13	3.31	0.80	0.09	4.15	0.92	0.06	3.44
oxalate/citrate × glycosaminoglycans ratio (mmol × 10^4)	19.09	7.64	34.80	43.17	7.42	257.85	195.49	7.72	631.77

[a] Md, median; min, minimum; max, maximum.

Table 2. Validity indexes for Examined Promoters and Inhibitors of Urolithiasis

parameters	Cut-off value		Accuracy	Specificity	Sensitivity
	Normal children	Urolithiasis			
urine saturation	< 4.70	≥ 4.70	88.89% (75.15-95.84%)	93.33% (71.27-99.67%)	86.67% (70.90-95.62%)
calcium/creatinine (mmol/mmol)	< 0.20	≥ 0.20	82.22% (67.42-91.49%)	73.33% (47.47-90.90%)	86.67% (70.90-95.62%)
oxalate/creatinine (mmol/mol)	< 48.00	≥ 48.00	75.56% (60.14-86-61%)	53.33% (26.68-76.80%)	86.67% (70.90-95.62%)
citrat/calcium ratio (mmol/mmol)	> 1.38	≤ 1.38	91.11% (77.87-97.11%)	73.33% (47.47-90.90%)	100% (90.05-100%)
magnesium/calcium × oxalate ratio (mmol)	a	a	a	a	a
oxalate/citrate × glycosaminoglycans ratio (mmol × 10^4)	10.16-34.80	< 10.16 or >34.80	97.78% (86.77-99.88%)	93.33% (71.27-99.67%)	100% (90.05-100%)

a - not possible to determine

Table 3. Number of positive risk factors in normal, hematuria and urolithiasis groups according to cut-off values

Groups	Number of Positive Risk Factors					
	0	1	2	3	4	5
Normal children	5 (33.33%)	4 (26.67%)	5 (33.33%)	1 (6.67%)	0	0
Hematuria	1 (2.78%)	5 (13.89%)	4 (11.11%)	9 (25.00%)	12 (33.33%)	5 (13.89%)
Urolithiasis	0	0	0	1 (3.33%)	10 (33.33%)	19 (63.33%)

MCR XVI. Mathematical Support for Combinatorial Chemistry

Dr. Bernhard Gruber
Institut für Organische Chemie und Biochemie
Technische Universität München
Lichtenbergstraße 4   *Tel.: ++49-89-289-13275*
D-85747 Garching   *Fax.: ++49-89-289-13294*
Germany   *EMail: gruber@c2h5oh.org.chemie.tu-muenchen.de*

Abstract

The algebra of the *s*- and *r*-vectors is an adequate formal tool to describe chemical objects in an abstract way. *Compounds* as well as *reactions* are represented including all constitutional and configurational aspects. The stereochemistry of simple organic molecules as well as of metal-organic compounds may be described in a unique way. Ionic bonds, covalent bonds, aromatics and electron deficiency compounds can be formally described without loss of information. Even reaction types and the flow of electrons can be described using this algebra. The biggest benefit of this approach is its intrinsic group theoretical structure. This does not bother the chemist for its use but allows the computer to handle and structure huge amounts of chemical data. This is especially important for combinatorial chemistry.

Usually, classical chemical syntheses from *n* starting materials require sequences of at least *n*–1 preparation steps including separation and purification of the intermediates. A perfect alternative for rapid syntheses of large varieties of agrochemically and pharmaceutically relevant products are one-pot syntheses by multicomponent reactions (MCR). Four to seven different types of participants (i. e. different isocyanides, amines, etc.) mixed in a reaction vessel undergo the transformation to one molecule. Using more than one representative of each type of starting materials, all possible combinations will lead to a molecular library of products formed according to the given reaction scheme. This is a preposition for finding compounds with desired properties.

The main efforts are to develop procedures that lead to the optimal compound with specific properties, i. e. methods *how* the optimal compound with best effects and least side effects can be found.

The design of library syntheses and the handling of the results require adequate mathematics and computer tools. For designing molecular libraries containing the sought compound one needs other information than for classical synthesis planning. How many different compounds will the library contain? How similar or different are these? Will the compounds show functional groups in similar spatial arrangements? The basic problem is the management of the flood of information which is generated. The combinatorial product space based on MCR approaches contains by magnitude more structures than all existing structure databases together. Using a combination of two Ugi-four-component reactions (involving di-carboxylic acids) the available product space will cover 10^{14} different structures from 500 different starting materials. Such vast numbers of data can not be handled by usual database systems. At present the automated syntheses of these compounds would take thousands of years.

Data may not be assigned to single structures but to *sets* of structures, determined by the starting compounds and the resulting reactions. Instead of comparing single structures one must compare collections without having the individual elements. Individuals can not be

Fig. 2: Two Ugi-4CR: one with a carboxylic acid, the other with water as an acid component

There are some requirements so that multicomponent chemistry works for combinatorial synthesis. Beside the typical essentials like yields and selectivity, the MCR must operate for a wide variety of representatives of the educt classes, i. e. the MCR must be *generic*, i. e. most of the starting materials will react according to the given reaction scheme. This is usually given, if the single reaction steps follow a certain procedure: Some of the reaction steps between starting materials and intermediates equilibrate, while the final steps, which proceed towards the desired product, are irreversible (Fig. 3). This type is called a type-II-MCR. Usually quantitative yields and an immensely large variability of starting materials and products result.[7]

Fig. 3: Type-II-MCR have a final quasi irreversible step

All the side-reactions must also be reversible. In the instance of the Ugi-4CR this is the case for the Hellmann-Opitz-3CR but not for the Passerini-3CR. The latter must be excluded by optimizing the reaction conditions towards the Ugi-4CR which in this case is easy to achieve by the solvent.

1.2 MCR in Combinatorial Practice

In 1961, when Ugi suggested to use MCR in a combinatorial way to produce molecular libraries[6], automated techniques for synthesis or screening were not developed yet. Then the paradigm in preparative organic chemistry was the syntheses of pure compounds. In the early nineties this paradigm shifted as there emerged a need for more compounds to match the improved and accelerated screening capabilities.

The combinatorial multicomponent chemistry gives a well-known degree of diversity and a high number of compounds.[4][8] The size of the molecular space M made up by a nCR is defined by the size of the n educt classes $n_1, n_2, ..., n_n$:

$$|B| = \prod_{i=1}^{n} n_i \qquad (1)$$

In fact the Ugi-4CR produces a stereo-centre at the carbonyl-C if non-symmetric aldehydes (formaldehyde) or ketones are used. Considering the available starting materials (mostly found in the catalogues of leading chemical manufacturers) there are about 10^{14} different combinations. This set of combinations is called the *Ugi-4CR product space*. Any molecular Ugi-4CR library is a subset of this product space.

Dr. Bernhard Gruber
Lehrstuhl OC1, Prof. Dr. Ivar Ugi
Institut für Organische Chemie und Biochemie
Technische Universität München

obtained by selection from an intractable set but by selective generation. Such highly efficient methods require a cleverly thought-out representation of the chemical objects *compound* and *reaction*, as it is given by the algebra of the *s*- and *r*-vectors.

Originally designed for describing the stereochemistry of chemical reactions via permutational isomerism, the algebra of the *s*- and *r*-vectors is useful to manage molecular libraries as well. It covers all combinatorial, constitutional, stereochemical and topological aspects of combinatorial chemistry on a profound mathematical basis. The little mathematics used, mostly group theory, will be explained and illustrated with chemical examples.

1 Principles of MCR-Based Combinatorial Chemistry

Combinatorial chemistry[1][2][3] is a rather simple technique to reach many different compounds within a short period of time. Basically there are two different approaches: One approach is the construction of chains of molecules with building blocks, usually all of the same kind, for example peptides out of amino acids. This must be done step-wise to get well-defined sequences. Multicomponent reactions (MCR), however, differ from this principle due to the construction plan that is inherent to them.[4] The sequence is always determined by the reaction scheme. No sophisticated procedures are needed. All combinations of the starting materials according to the reaction scheme are possible (Fig. 1)The set of combinations is called a *molecular library*.
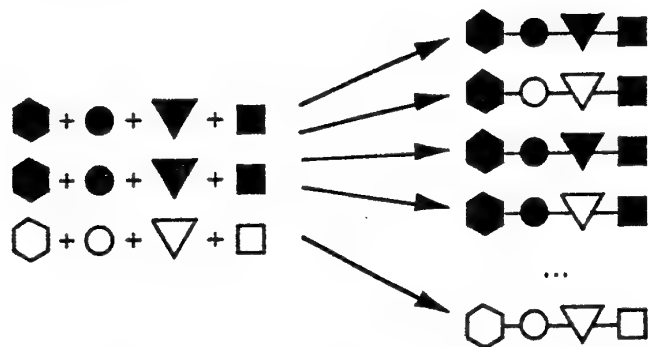
Fig. 1: Scheme of the combinatorial construction principle of a molecular library using a four-component reaction

Using combinatorial methods also means that you obtain a lot of information within a short period of time. In order to reduce the costs for information management computers should be helpful.

1.1 Multicomponent Reactions for Combinatorial Chemistry

From the methodological point of view MCR are the kind of chemistry that is best fitting for combinatorial synthesis: MCR may be carried out on solid or in liquid phase by so-called "Eintopf-Reaktionen" (one-pot reactions). In the latter approach, the starting materials are simply mixed and the program for combining the starting materials is given by the chemistry of the MCR. For example, the Ugi-four-component-reaction (Ugi-4CR) (Fig. 2) is combining an amine, an aldehyde (or a ketone), a carboxylic acid (or other proton donors) and an isocyanide.[5][6] These are called the *educt classes*.

Dr. Bernhard Gruber
Lehrstuhl OC1, Prof. Dr. Ivar Ugi
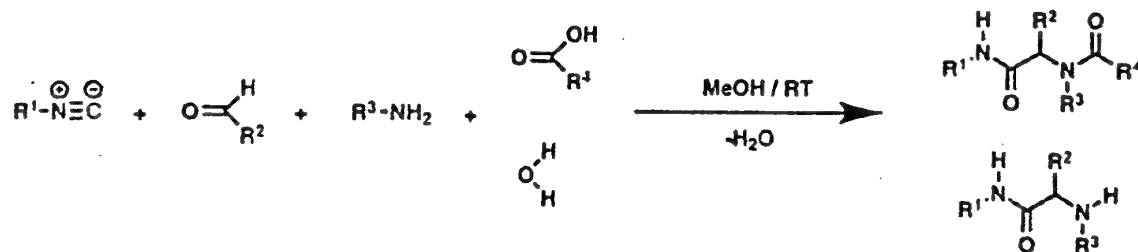Institut für Organische Chemie und Biochemie
Technische Universität München

Fig. 2: Two Ugi-4CR: one with a carboxylic acid, the other with water as an acid component

There are some requirements so that multicomponent chemistry works for combinatorial synthesis. Beside the typical essentials like yields and selectivity, the MCR must operate for a wide variety of representatives of the educt classes, i. e. the MCR must be *generic*, i. e. most of the starting materials will react according to the given reaction scheme. This is usually given, if the single reaction steps follow a certain procedure: Some of the reaction steps between starting materials and intermediates equilibrate, while the final steps, which proceed towards the desired product, are irreversible (Fig. 3). This type is called a type-II-MCR. Usually quantitative yields and an immensely large variability of starting materials and products result.[7]

a $\xleftarrow{\quad + b \quad}$ **a-b** $\xleftrightarrow{\quad}$ **a-b** c $\xrightarrow{\quad + d \quad}$ **a-b** c**-d**

Fig. 3: Type-II-MCR have a final quasi irreversible step

All the side-reactions must also be reversible. In the instance of the Ugi-4CR this is the case for the Hellmann-Opitz-3CR but not for the Passerini-3CR. The latter must be excluded by optimizing the reaction conditions towards the Ugi-4CR which in this case is easy to achieve by the solvent.

1.2 MCR in Combinatorial Practice

In 1961, when Ugi suggested to use MCR in a combinatorial way to produce molecular libraries[6], automated techniques for synthesis or screening were not developed yet. Then the paradigm in preparative organic chemistry was the syntheses of pure compounds. In the early nineties this paradigm shifted as there emerged a need for more compounds to match the improved and accelerated screening capabilities.

The combinatorial multicomponent chemistry gives a well-known degree of diversity and a high number of compounds.[4][8] The size of the molecular space M made up by a nCR is defined by the size of the n educt classes $n_1, n_2, ..., n_n$:

$$|B| = \prod_{i=1}^{n} n_i \tag{1}$$

In fact the Ugi-4CR produces a stereo-centre at the carbonyl-C if non-symmetric aldehydes (formaldehyde) or ketones are used. Considering the available starting materials (mostly found in the catalogues of leading chemical manufacturers) there are about 10^{14} different combinations. This set of combinations is called the *Ugi-4CR product space*. Any molecular Ugi-4CR library is a subset of this product space.

Dr. Bernhard Gruber
Lehrstuhl OC1, Prof. Dr. Ivar Ugi
Institut für Organische Chemie und Biochemie
Technische Universität München

TLM

A very suitable device for the automated synthesis of MCR libraries are micro titre plates (Fig. 4). These are available in different densities up to some 3000 wells per plate. Figure 4 shows how a 8×12 micro titre plate can be used for the synthesis of a 96-library by mixing 4 aldehydes, 3 amines, 4 carboxylic acids and 2 isocyanides. Each well contains an unique mixture from which an Ugi-4CR product may evolve. This is defined by a 2-dimensional location. The mixture (A2, B1, C1, D2) will be positioned at (4, 5).
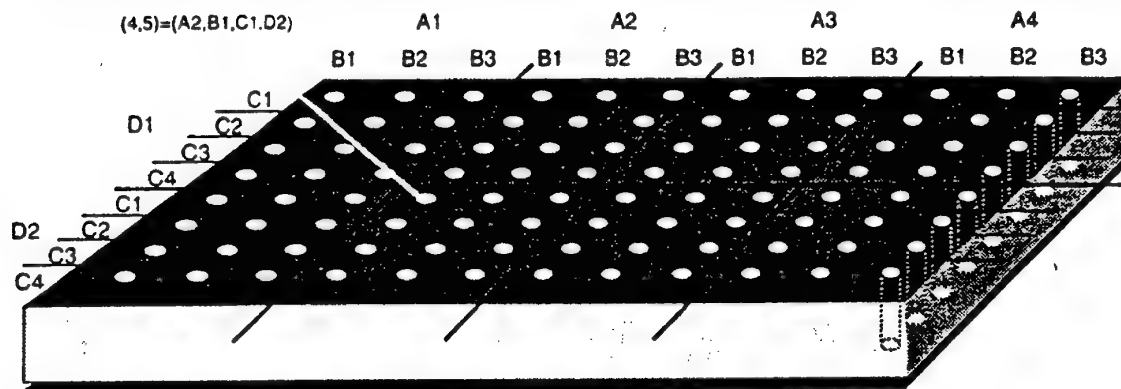
Fig. 4: Distribution of 13 different starting materials on a micro titre plate in order to produce a 96-library.

Figure 5 shows an array of micro titre plates necessary for the synthesis of a 960'000-library emerging from 40 aldehydes, 30 amines, 40 carboxylic acids and 20 isocyanides. This is still far away from the possibilities of the Ugi-4CR.
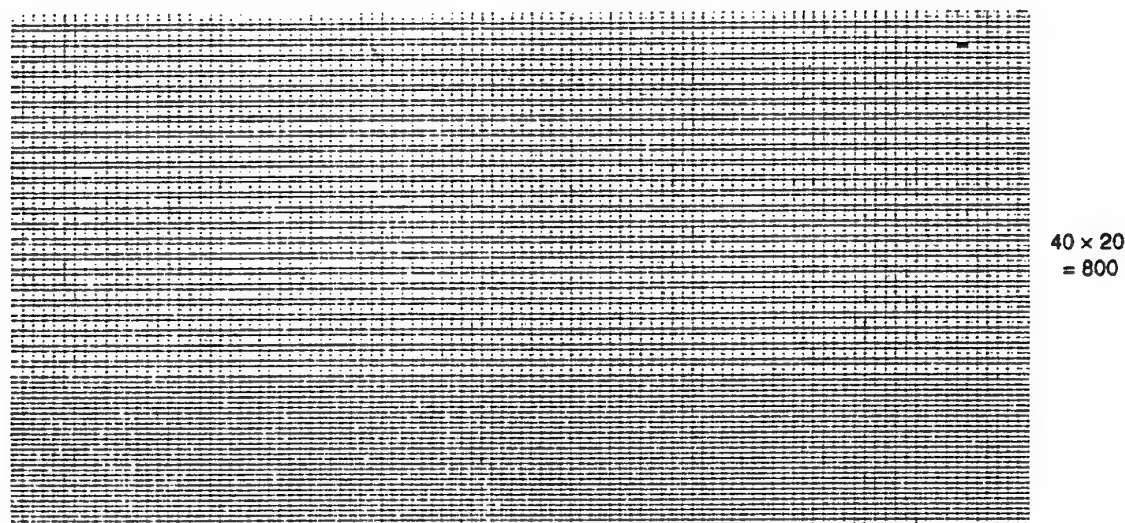
Fig. 5: An array of micro titre plates necessary for a 960'000 library.

To use the high density titre plates is not the solution of the problem because this reduces the problem only by a factor of 10^2. To give an impression of the size of the problem: Assume each well will be filled with 1µg starting materials. Further assume that —at the present state of the art— an automaton fills each titre plate within 1 second. Then, the complete Ugi-4CR product space with 10^{14} different combinations will consume 100 tons of starting materials within 33'000 years.

Dr. Bernhard Gruber
Lehrstuhl OC1, Prof. Dr. Ivar Ugi
Institut für Organische Chemie und Biochemie
Technische Universität München

1.3 Need for Computers

To find the best matching compound concerning sought properties out of the available product spaces is an optimization problem, because there is no algorithm mapping properties onto a (set of) compound(s). For the designer of a molecular library all intractability of this problem results out of the size of the product spaces. No matter if selecting starting materials corresponding with a given specification, or specifying a sublibrary, or optimizing a lead, one always has to face the enormous molecular space one works within. It is not possible to produce the whole molecular space and then select the interesting structures. No computer could store nor investigate that much information. The information must be generated selectively. Answers must be found on the level of the starting compounds and construction principles that are given by the chosen multicomponent chemistry.

Effective and efficient representations of the chemical objects are needed for computer support. As so often in chemistry, group theory helps a lot.

1.4 Formal Basis

There is little mathematics needed for a formal representation of combinatorial chemistry. The complete mathematical discourse on this topic is given in the mentioned literature. This chapter gives all necessary information without proofs but explains the most important mathematical correlations.

1.4.1 Permutations

A *permutation* $\pi = (l_a\ l_b\ l_c\ ...\ l_s)$ is a mapping from the set $L \supseteq \{l_a\ l_b\ l_c\ ...\ l_s\}$ onto L.

$$\pi: L \to L$$
$$\pi\ \ bijective$$
(2)

according to $\pi\ (l_a) = l_b, \pi\ (l_b) = l_c, ..., \pi\ (l_s) = l_a$. All further elements are mapped to themselves. This mapping may also be noted as a vector

$$\pi = \begin{bmatrix} l_a & l_b & ... & l_s \\ l_b & l_c & ... & l_a \end{bmatrix}$$
(3)

A composition of two permutations $\pi_2 \circ \pi_1 = \pi_2\pi_1$ applies first π_1 and then π_2. If two permutations are *disjoint* (that means they have no elements in common) the sequence of application has no effect. i. e. $(1\ 5)(3\ 4) = (3\ 4)(1\ 5)$. Also the element a cycle starts with makes no difference. i. e. $(1\ 5\ 6) = (5\ 6\ 1)$. but naturally the sequence *within* the cycle does, i. e. $(1\ 5\ 6) \neq (1\ 6\ 5)$! Non-disjoint permutations can be "normalized" so that the cycles become disjoint. i. e. $(1\ 5)(1\ 5\ 6)$ can be transformed to $(5\ 6)$.

1.4.2 Automorphisms and Groups for Structuring Chemical Information

A *group* $G = (L, \cdot)$ is a pair of a set L and an operation $\cdot$ in L with the following properties:

   (1) *Closure*: For any two elements $\alpha, \beta \in L$, the product $\alpha \cdot \beta$ is also element of L;

   (2) *Associativity*: For all elements $\alpha, \beta, \gamma \in L$, we have $\alpha \cdot (\beta \cdot \gamma) = (\alpha \cdot \beta) \cdot \gamma$;

   (3) *Existence of an identity element*: There exists an element $e \in L$ such that for all ele-

ments $\alpha \in L$ holds $\alpha \cdot \varepsilon = \varepsilon \cdot \alpha = \alpha$;

(4) *Existence of inverse elements*: For any element $\alpha \in L$ there exists an element α^{-1} such that $\alpha \cdot \alpha^{-1} = \alpha^{-1} \cdot \alpha = \varepsilon$.

This paper is exclusively talking about finite groups, i. e. groups with a finite set of elements.

Each $U = (M, \cdot)$, $M \subseteq L$, is called a *subgroup* of G, written $U < G$, if U is closed concerning the operation $\cdot$. Groups (and subgroups) are usually defined by a set of *generators S*. All possible combinations of elements of set S will give the group $<S>$.

For example, the set of permutations $L = \{(), (1\ 2\ 3), (1\ 3\ 2), (1\ 2)(4\ 5), (1\ 3)(4\ 5), (2\ 3)(4\ 5)\}$ with the composition of permutations $\circ$ forms a group $G = (L, \circ)$. The set $M = \{(), (1\ 2\ 3), (1\ 3\ 2)\}$ forms a subgroup $U = (M, \circ) < G$. The set $S = \{(1\ 2), (1\ 2\ 3\ 4\ 5)\}$ generates the group $\mathfrak{S}_5$ that covers all 120 possible permutations of up to five elements. G as well as U are subgroups of $\mathfrak{S}_5$.

The group-theoretic notion *coset* plays an important role. A *coset space* of G is generated by a subgroup $U < G$. For any $\alpha \in G$ the set $\alpha \cdot U = \alpha U$ is called a *left-coset* of U in G. The left-coset αU can be described as $\alpha U = \{\gamma \in G: \alpha^{-1}\gamma \in U\}$. In section 2 will be shown that each coset represents exactly one product of a library. The *index* $[G:U]$ of U in G is the number of distinct (left-)cosets. All the left-cosets of U in G have the same size. Because the identity element is member of G, U itself is a left-coset. Therefore all cosets of U in G have the same size as U. Each $\gamma \in G$ belongs to at least one left-coset of U in G and any two left-cosets of U in G are either identical or disjoint. Therefore G is a disjoint union of all the cosets of U in G. The index $[G:U]$ is equal to $|G| / |U|$. Each coset is uniquely represented by any of its members together with the subgroup U. A set of *one* representative per coset is called a *travese*.

Groups are capable of structuring sets (of molecules) by making up a coset space. In the example above two permutations define a set of 120, a small subset that generate a group structures these 120 elements into cosets. This is why group theory is of high value for structuring big sets.

This approach has been initiated by Dugundji and Ugi in 1984 and refined by Dietz, Gruber and Ugi.[11][12][13][14] Several kinds of equivalency relations (for example *chemical identity*) have been applied to structure chemical information.

1.4.3 Automorphisms for Molecules and Reactions

Groups based on permutations are called *automorphism groups*. Automorphism groups are extremely useful, not only to describe stereochemistry of molecules, but also for constitutional phenomena, so-called chemical reactions. In 1971 Ugi and Dugundji represented the constitutional part of chemistry by redistribution of binding electrons formalized as the *algebra of the* be- *and* r-*matrices*.[15] In 1992 an algebraic model of stereochemistry was given, involving stereochemistry into chemical reactions and vice versa. Section 3.7 of [16] is dealing with the fusion of the theory of the chemical identity group with the algebra of the *be-* and *r*-matrices. The result was the *algebra of* s- *and* r-*vectors*.[12][13][16][17] s stands for *stereochemical* and r for *reaction*. The derivation of this algebra is also given in Ref. [13]pp9-11. The algebra of s- and r-vectors is capable of describing structures and reactions regarding stereochemistry [13], delocalized electron systems and electron deficient compounds [17].

Each atom has located topological positions that function as connection points. A sp^3-carbon possesses typically four topological positions. The positions are called *topological* but not *geometric* because their geometric location is not known but their neighbourhood is.

Chemical structures are represented by one-to-one mappings of the topological positions of (different) atoms. This is isomorphic to permuting a finite set of ligands, but unlike the theory of the chemical identity group there is no distinction between atoms of the ligands and atoms of the skeleton any more. Any (set of) molecule(s) is represented by a permutation. Therefore chemical reactions are simply permutations applied onto permutations.

Fig. 6 shows the constitutional changes of the Ugi-4CR which is represented by the matrix R. The matrix E represents the starting materials: an aldehyde, an amine, a carboxylic acid and an isocyanide. The Ugi-product and one molecule of water are represented by the matrix B.

Matrix E:

	Ra	C	O	O	H	Rb	N	H	H	Rc	C	H	O	Rd	N	C
Ra		1														
C	1		2	1												
O		2	4													
O		1		4	1											
H				1												
Rb							1									
N						1	2	1	1							
H							1									
H							1									
Rc											1					
C										1		1	2			
H											1					
O											2		4			
Rd															1	
N														1	2	2
C															2	2

E

$+$

Matrix R:

	Ra	C	O	O	H	Rb	N	H	H	Rc	C	H	O	Rd	N	C
Ra																
C			-1		1											
O																
O	-1			-1								2				
H			-1							1						
Rb																
N	1						-1	-1	1							
H							-1					1				
H							-1					1				
Rc																
C							1					-2			1	
H																
O			1					1	-2							
Rd																
N							1									-2
C		2						1					-2	-2		

R

$=$

Matrix B:

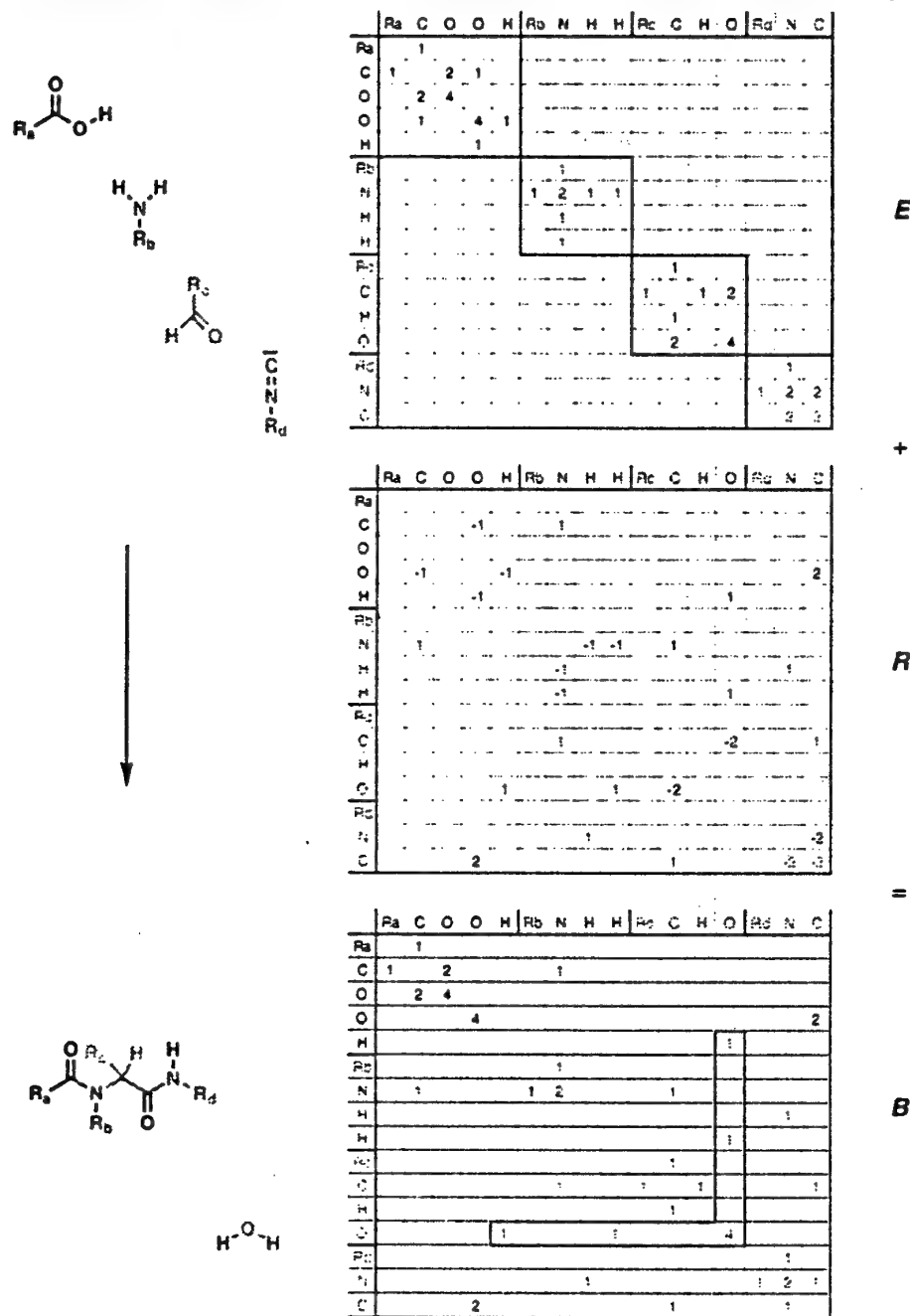|	Ra	C	O	O	H	Rb	N	H	H	Rc	C	H	O	Rd	N	C
Ra	1															
C	1		2				1									
O		2	4													
O			4												2	
H												1				
Rb							1									
N	1					1	2			1						
H												1				
H												1				
Rc											1					
C									1	1					1	
H											1					
O				1			1					4				
Rd															1	
N							1							1	2	1
C		2							1						1	

B

Fig. 6: Ugi-4CR represented as *be-&r*-matrices

The equation $B = R+E$ is fundamental to the algebra of the *be*-and *r*-matrices. The matrices exclusively have information about the distribution of the binding electrons, whereas the *r*- and *s*-vectors are representations of the topological positions on the atoms (Fig. 7).[12] The first line is needed as a reference what topological positions are mapped. The *s*-vector ε is defining the starting materials. ρ is the Ugi-4CR *r*-vector. The *s*-vector β results from the application of ρ onto ε: $\beta = \rho \circ \varepsilon$. Fig. 7 shows the Ugi-4CR with the topological positions and two different notations for the permutations ε, ρ and β: as vectors and as cycles (s. section 1.4.1).

Fig. 7: The Ugi-4CR noted in terms of *s*- and *r*-vectors

Molecules (or ensembles of them) correspond with permutations. Families of permutational isomers correspond with automorphism groups. The algebra of *s*- and *r*-vectors allows to expand this approach onto sets of "isomeric" ensembles of molecules.

2 A Precise Representation for Molecular Libraries

As well as automorphism groups are useful for representing families of permutational isomers in an efficient way, they can represent molecular libraries.[16][19] Molecular libraries are used in combinatorial chemistry for finding and optimizing new drugs. There is a big space of molecules corresponding with each MCR. Libraries are subsets of this space. In order to decide *what* libraries are most expecting one must use tools that determine these sets.

Due to its size the entire molecular space of an MCR can not be produced as compounds, nor can it be represented as a list. Group-theory is capable of representing a set of molecules by a subset (a so-called set of generators) and a construction principle. This saves memory and reduces the exponential complexity to quadratic complexity, which is important for both computers and the designer of molecular libraries.

Notions of group theory have chemical interpretations. For example a group may define an equivalence relation. That means the members of the group are *equivalent* concerning a defined property, for example they belong to the same aromatic system. Subgroups of this group may refine the equivalence relation. As for example the subgroup given by all permutations applied onto molecular parts with a maximum mass of 150 daltons.

2.1 Molecular Libraries represented by Notions of Group Theory

Each *n*-CR leads to a specific backbone with *n* topological sites. At these sites different ligands will be positioned by different starting materials. Each kind of starting material has its well-

defined site due to the "deterministic" reaction mechanisms. The amines in the Ugi-4CR (Fig. 2) will always define the same part of the backbone, etc. By transposing the subgraph defined by the amine with another amine you get a different "permutational isomer" which is a member of the library. The only difference between families of permutational isomers and molecular libraries is that in the latter are permutations applied upon *extra-molecular* parts.

Fig. 7 shows the Ugi-4CR in terms of s- and r-vectors. The Ugi-4CR (as well as any other 4CR of type II) can be simplified to the scheme given in Fig. 3: The result of the Ugi-4CR of educts a, b, c and d will be a molecule a—b—c—d. The educts must be of the appropriate classes A, B, C, D, i. e. "a" must be an acid component (class A), "b" an amine (class B), "c" an aldehyde or ketone (class C) and "d" an isocyanide (class D).

The size of the product space is determined by the size of the educt classes $|A|$, $|B|$, $|C|$ and $|D|$. Assume that the commercially available number of educts are like $|A| = 200$, $|B| = 200$, $|C| = 150$ and $|D| = 20$. According to formula (1) the resulting product space M will cover $|M| = |A| \cdot |B| \cdot |C| \cdot |D| = 200 \cdot 200 \cdot 150 \cdot 20 = 1.2 \cdot 10^8$ different products.

For a construction principle for Ugi-4CR libraries, let the members of the educt classes be identified by $a_1, ..., a_{200}, b_1, ..., b_{200}, c_1, ..., c_{150}, d_1, ..., d_{20}$. Furthermore let the sequence $m = a_1$—b_1—c_1—d_1 represent a specific Ugi-4CR reference product. Then $s = \{(a_1\ a_2),$ $(a_1\ a_2 ... a_{200}), (b_1\ b_2), (b_1\ b_2 ... b_{200}), (c_1\ c_2), (c_1\ c_2 ... c_{150}), (d_1\ d_2), (d_1\ d_2 ... d_{20})\}$ is a set of generators of the group $S = <s>$, that covers all permutations concerning members of the same class. Such permutations are for example $\alpha = (a_1\ a_2)$, as well as $\beta = (a_1\ a_{50})(b_1\ b_2)$ and $\gamma = (a_2\ a_3)$, but not $\pi = (a_1\ b_1)$ nor $\rho = (a_1\ b_2)(b_1\ a_2)$ because the molecules $\alpha \circ m = a_2$—b_1—c_1—d_1, $\beta \circ m = a_{50}$—b_2—c_1—d_1 as well as $\gamma \circ m = a_1$—b_1—c_1—d_1 ($= m$) are valid in the sense that the position "a" is always kept by a member of class A, "b" by B and so on. This is definitely not true for $\pi \circ m = b_1$—a_1—c_1—d_1 nor for $\rho \circ m = b_2$—a_2—c_1—d_1.

As γ shows, there are permutations that are valid but do not affect the molecule. These permutations transpose members "within" the class itself. They are defined by the group $T = <\{(a_2\ a_3), (a_2\ a_3 ... a_{200}), (b_2\ b_3), (b_2\ b_3 ... b_{200}), (c_2\ c_3), (c_2\ c_3 ... c_{150}), (d_2\ d_3),$ $(d_2\ d_3 ... d_{20})\}>$, i. e. the elements a_1, b_1, c_1 and d_1 are not involved.

Each left-coset $\alpha \circ T$ of T in S corresponds with exactly one molecule of the library. This means the library is defined by the left coset space $M = \{\alpha T \mid \alpha \in S\}$. The size of M is given by

$|M| = |S| / |T|$
$= (|A|! \cdot |B|! \cdot |C|! \cdot |D|!) / ((|A|-1)! \cdot (|B|-1)! \cdot (|C|-1)! \cdot (|D|-1)!)$
$= |A| \cdot |B| \cdot |C| \cdot |D|$
$= 1.2 \cdot 10^8$

A molecular library is defined by the coset space or any corresponding traverse (section 1.4.2).

2.2 Costs

The computer does not calculate all the elements of a group. Instead a so-called representation matrix is built.[16][18] This is an upper triangle matrix of size $n \times n$ where n is the number of elements to permute, in this case the number of educts. The process to construct a representation matrix runs with quadratic time- and space-complexity respective the number n of elements to permute.

Dr. Bernhard Gruber
Lehrstuhl OC1, Prof. Dr. Ivar Ugi
Institut für Organische Chemie und Biochemie
Technische Universität München

In the given example of the Ugi-4CR product space there are $|A|+|B|+|C|+|D| = 570$ educts to distribute on the backbone which corresponds with a subset of $\mathfrak{S}_{570}$, the set of all possible permutations on 570 objects.

To store $\mathfrak{S}_{570}$ as a representation matrix you need 162.165 permutations instead of 570! ($\approx 10^{1000}$). The representation matrix of $S < \mathfrak{S}_{570}$ has 51.140 defined entries and $T < S$ is only 50.599. Using this concept one can do membership tests with quadratic complexity and can easily construct subgroups of already represented groups.

3 Applications of this Approach

In order to give an idea how the representation matrices look like and how the concept may be applied, a rather small example is given. It bases upon an Ugi-4CR library each class having 5 members (5·5·5·5-Ugi-4CR library). Bigger examples can easily be handled by computers but are not of the right size to be printed.

3.1 The Management of a 5·5·5·5-Ugi-4CR Library

The following example uses the Ugi-4CR with 5 compounds of each class. The 5 carboxylic acids are responsible for filling the part R_a of the Ugi-4CR-backbone (Fig. 7) with the molecular rests 1, 2, 3, 4 and 17. The numbers are arbitrary but unique. The 5 aldehydes are filling part R_b of the Ugi-4CR-backbone with the molecular rests 5, 6, 7, 8 and 18. The 5 amines are filling part R_c of the Ugi-4CR-backbone with the molecular rests 9, 10, 11, 12 and 19. And the 5 isocyanides are filling part R_d of the Ugi-4CR-backbone with the molecular rests 13, 14, 15, 16 and 20. One expects $5^4 = 625$ different products.

Each of the resulting Ugi-products is represented by a sequence a—b—c—d, where $a \in \{1, 2, 3, 4, 17\}$, $b \in \{5, 6, 7, 8, 18\}$, $c \in \{9, 10, 11, 12, 19\}$ and $d \in \{13, 14, 15, 16, 20\}$. One of the products is given by the sequence 17—18—19—20. This shall be the reference product. The other products can be received by replacing 17 by 1, 2, 3 or 4, replacing 18 by 5, 6, 7 or 8, etc. The resulting set of permutations is $g = \{(17\ 4), (17\ 3), (17\ 2), (17\ 1), (18\ 8), (18\ 7), (18\ 6), (18\ 5), (19\ 12), (19\ 11), (19\ 10), (19\ 9), (20\ 16), (20\ 15), (20\ 14), (20\ 13)\}$ and all their combinations.

By combining the permutations of set g you can describe all the different products of the library. Unnecessarily, many of the permutations represent the same product, because some of the combinations result in permutations that do not affect the reference product. As for example the permutation $(1\ 2)$ that results from the combination $(1\ 17) \circ (2\ 17) \circ (1\ 17)$. These duplicates may be handled mathematically: The set of all combinations of permutations of set g forms the group $G = \langle g \rangle$. G has $5!^4 = 207'360'000$ members. The subgroup $U < G$ that results from all possible combinations that do *not* affect the reference isomer is generated by $U = \langle\{(1\ 2), (2\ 3), (3\ 4), (5\ 6), (6\ 7), (7\ 8), (9\ 10), (10\ 11), (11\ 12), (13\ 14), (14\ 15), (15\ 16)\}\rangle$. U has $4!^4 = 331'776$ members.

In the above example the index of U in G is $|G| / |U| = 5!^4 / 4!^4 = 5^4 = 625$. The left-coset space of U in G is the family of sets $\{\gamma U \mid \gamma \in G\}$. For any $\gamma \in G$ the left-coset γU describes exactly one member of the library. For example the left-coset $(1\ 17)(5\ 6)(12\ 19)U$ contains the permutation $(1\ 17)(12\ 19)$, so it represents the product 1—18—12—20. There is no other product represented by this left-coset and no other left-coset representing the product. Any traverse of the coset space defines the complete library. Naturally not the traverse is calculated but the representation matrices of G and U (Figures 8 and 9).

Dr. Bernhard Gruber
Lehrstuhl OC1, Prof. Dr. Ivar Ugi
Institut für Organische Chemie und Biochemie
Technische Universität München

3.2 Program behaviour

The group G is generated by the set of permutations $g = \{(17\,4), (17\,3), (17\,2), (17\,1), (18\,8), (18\,7), (18\,6), (18\,5), (19\,12), (19\,11), (19\,10), (19\,9), (20\,16), (20\,15), (20\,14), (20\,13)\}$. The representation matrix of group G is of dimension 20:

	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	0
2																			0	(2 3 4 17 1)
3																		0	(3 4 17 2)	(3 4 17 1)
4																	0	(4 17 3)	(4 17 2)	(4 17 1)
5																0	undefined	undefined	undefined	undefined
6															0	(6 7 8 18 5)	undefined	undefined	undefined	undefined
7														0	(7 8 18 6)	(7 8 18 5)	undefined	undefined	undefined	undefined
8													0	(8 18 7)	(8 18 6)	(8 18 5)	undefined	undefined	undefined	undefined
9												0	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
10											0	(10 11 12 19 9)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
11										0	(11 12 19 10)	(11 12 19 9)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
12									0	(12 19 11)	(12 19 10)	(12 19 9)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
13								0	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
14							0	(14 15 16 20 13)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
15						0	(15 16 20 14)	(15 16 20 13)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
16					0	(16 20 15)	(16 20 14)	(16 20 13)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
17				0	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	(17 4)	(17 3)	(17 2)	(17 1)
18			0	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	(18 8)	(18 7)	(18 6)	(18 5)	undefined	undefined	undefined	undefined
19		0	undefined	undefined	undefined	undefined	undefined	undefined	(19 12)	(19 11)	(19 10)	(19 9)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
0	0	undefined	undefined	undefined	(20 16)	(20 15)	(20 14)	(20 13)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined

Fig. 8: Group G generated by 16 permutations contains all permutations corresponding with the library

The members of the group are found by composing the permutations of the representation matrix, taking exactly one permutation of each row. Each combination will give an unique member. Consequently, the size of the group can be calculated by multiplying the numbers of

defined matrix entries per row: $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot = 5!^4 = 207'360'000$. The Pascal-implementation on an Apple Macintosh 8500/180 needs less than one second to calculate the representation matrix out of the set of generators g.

0	(2 1)	(3 2 1)	(4 3 2 1)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
2	0	(3 2)	(4 3 2)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
3		0	(4 3)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
4			0	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
5				()	(6 5)	(7 6 5)	(8 7 6 5)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
6					()	(7 6)	(8 7 6)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
7						()	(8 7)	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
8							()	undefined	undefined	undefined	undefined	undefined	undefined	undefined	undefined
9								()	(10 9)	(11 10 9)	(12 11 10 9)	undefined	undefined	undefined	undefined
10									()	(11 10)	(12 11 10)	undefined	undefined	undefined	undefined
11										()	(12 11)	undefined	undefined	undefined	undefined
12											()	undefined	undefined	undefined	undefined
13												()	(14 13)	(15 14 13)	(16 15 14 13)
14													()	(15 14)	(16 15 14)
15														()	(16 15)
16	2	3	4	5	6	7	8	9	10	11	12	13	14	15	0

Fig. 9: Group U generated by 12 permutations

The set of generators of group $U<G$ is {(2 1), (3 2), (4 3), (6 5), (7 6), (8 7), (10 9), (11 10), (12 11), (14 13), (15 14), (16 15)}. The corresponding representation matrix is of dimension 16. The group U has $4 \cdot 3 \cdot 2 \cdot 1 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot = 4!^4 = 331'776$ members.

4 Discussion

Molecules (or ensembles of them) correspond with permutations. Families of permutational isomers correspond with automorphism groups. The algebra of s- and r-vectors allows to expand this approach onto sets of "isomeric" ensembles of molecules. In this sense MCR libraries are isomeric to the set of educts. Thereby automorphism groups are useful for representing families of permutational isomers as well as molecular libraries or any other structured sets of molecules in an efficient way.

The structuring properties of group theory are useful for the efficient storage of chemical data. The approach works for managing the data generated by combinatorial chemistry. Chemical properties correspond with group-theoretic structures like cosets or subgroups. This structures sets of molecules in a hierarchical manner. Thereby automorphism groups are useful for representing families of permutational isomers as well as molecular libraries or any other structured sets of objects in an efficient way.

5 References

(1) Gordon, E.M.; Barrett, R.W.; Dower, W.J.; Fodor, S.P.A.; Gallop, M.A.: "Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries". *J. Med. Chem.* **1994,**

Dr. Bernhard Gruber
Lehrstuhl OC1, Prof. Dr. Ivar Ugi
Institut für Organische Chemie und Biochemie
Technische Universität München

TUM

37, 1233-1251.

(2) Gallop, M.A.; Barrett, R.W.; Dower, W.J.; Fodor, S.P.A.; Gordon, E.M.: "Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions". *J. Med. Chem.* **1994**, *37*, 1385-1401.

(3) Balkenhohl, F.; von dem Bussche-Hünnefeld, C.; Lansky, A., Zechel, C.: "Combinatorial Synthesis of Small Organic Molecules". *Angew. Chem. Int. Ed. Engl.* **1996**, *35*, 2289-2337.

(4) Ugi, I.; Dömling, A., Gruber, B.; Almstetter, M.: "Multicomponent Reactions and Their Libraries - A New Approach to Preparative Organic Chemistry" *Croatica Chemica Acta* **1997**, *70*, 631-647.

(5) Ugi, I. et al.: "Versuche mit Isonitrilen" *Angew. Chem.* **1959**, *71*, 386.

(6) Ugi, I.; Steinbrückner, C.: "Isonitrile II. Reaktion von Isonitrilen mit Carbonylverbindungen, Aminen und Stickstoffwasserstoffsäure" *Chem. Ber.* **1961**, *94*, 734-742.

(7) Ugi, I. K. (1997) Croatica Chemica Acta, submitted.

(8) Ugi, I.; Goebel, M.; Gruber, B.; Heilingbrunner, M.; Heiß, C.; Hörl, W.; Kern, O.; Starnecker, M.; Dömling, A.: "Molecular Libraries in Liquid Phase via Ugi-MCR" *Res. Chem. Intermed.* **1996**, *22*, 625-644.

(9) Ugi, I.; Marquarding, D.; Klusacek, H.; Gokel, G.; Gillespie, P.: "Chemie und logische Strukturen" *Angew. Chem.* **1970**, *82*, 741-771.

(10)Ugi, I.; Dugundji, J.; Kopp, R.; Marquarding, D.: "Perspectives in Theoretical Stereochemistry" Lecture Notes in Chemistry, Springer, Berlin, Heidelberg, New York, Tokyo 1984.

(11)Dietz, A.; Gruber, B.; Ugi, I.: "Algorithmische Behandlung stereochemischer Problemstellungen" *match* **1994**, *31*, 37-87.

(12)Gruber, B.; Almstetter, M.; Heilingbrunner, M.: "MCR IV: Formalisierung chemischer Objekte für die Kombinatorische Chemie", in: Bernhard Koppenhoefer, ed. "Chemie und Informatik: Terra incognita oder das Land der unbegrenzten Möglichkeiten", Shaker-Verlag, Aachen, 1997.

(13)Ugi, I.; Bauer, J.; Blomberger, C.; Brandt, J; Dietz, A.; Fontain, E.; Gruber, B.; Scholley-Pfab, A. v.; Senff, A.; Stein, N.: "Models, Concepts, Theories, and Formal Languages in Chemistry and Their Use as a Basic for Computer Assistance in Chemistry" *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 3-16.

(14)Ugi, I.; Bauer, J.; Bley, K.; Dengler, A.; Dietz, A.; Fontain, E.; Gruber, B.; Herges, R.; Knauer, M; Reitsam, K.; Stein, N.: "Computer-Assisted Solution of Chemical Problems - The Historical Development and the Present State of the Art of a New Discipline of Chemistry" *Angew. Chem. Int. Ed.* **1993**, *32*, 201-227. "Computerunterstützte direkte Lösung chemischer Probleme - die Entstehungsgeschichte und der gegenwärtige Status einer neuen Disziplin der Chemie" *Angew. Chem.* **1993**, *105*, 210-239.

(15)Dugundji, J.; Ugi, I.: "An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs" *Top. Curr. Chem.* **1973**, *39*, 19-64.

(16)Gruber, B.: "Algebraische Modellierung der Stereochemie" Dissertation, Technische Universität München, München 1992.

(17)Gruber, B.: "Eine lineare algebraische Repräsentation für Objekte der Syntheseplanung" in R. Moll (Ed.), Software-Entwicklung in der Chemie 9, Proceedings of the 9th Workshop "Computer in Chemistry" 16.-18. November 1994, Bitterfeld, 99-111, Gesellschaft Deutscher Chemiker, Frankfurt 1995.

(18)Hoffmann, C.M.: *Group-Theoretic Algorithms and Graph Isomorphisms.* Lecture Notes in Computer Science 136, Springer, Berlin-Heidelberg-New York 1982. Algorithms are available via http://c2h5oh.org.chemie.tu-muenchen.de/persons/gruber/algorithmen/.

(19)Gruber, B.: "MCR XV. Combinatorial Chemistry can be Formalized!". German-Polish Workshop on Multicomponent Reactions & Combinatorial Chemistry,Rzeszow 28.-30.09.1997, Proceedings, ed. Zdzislaw S. Hippe.

Dr. Bernhard Gruber
Lehrstuhl OC1, Prof. Dr. Ivar Ugi
Institut für Organische Chemie und Biochemie
Technische Universität München

Fig. 1

Fig. 2

a $\xrightarrow{\text{+ b}}$ a-b $\longleftrightarrow\longrightarrow$ a-b c $\xrightarrow{\text{+ d}}$ a-b c-d

Fig. 3

Fig. 4

(4,5)=(A2,B1,C1,D2)

Fig. 4
100%

MacDraw Pro 1.5
Verleger

Gruber:...:sp3
Gruber:...:Tetraeder

40 × 30 = 1200

40 × 20
= 800

Fig. 5

Fig. 6

E
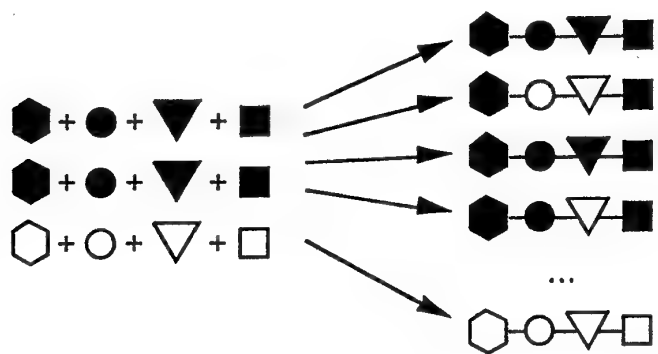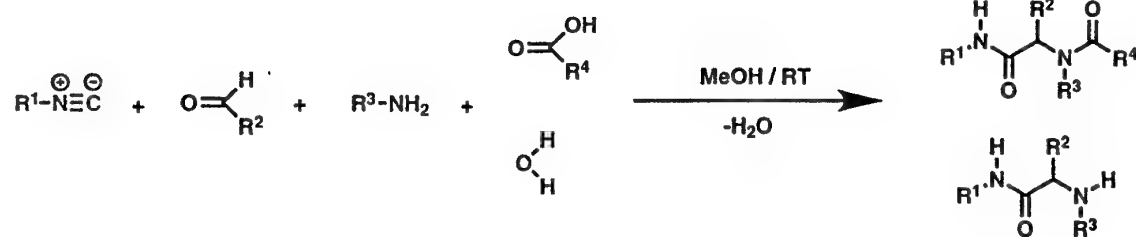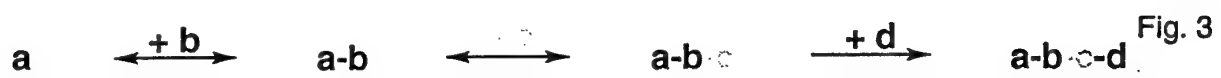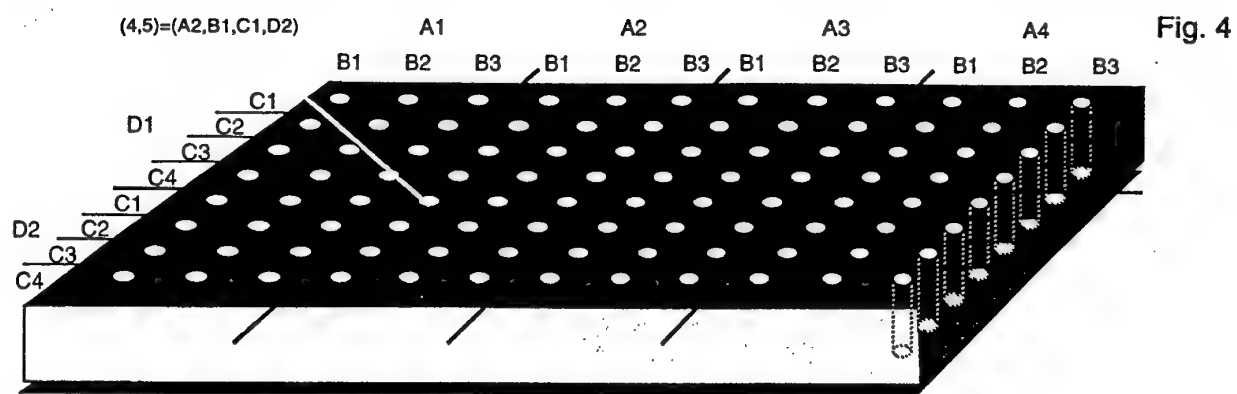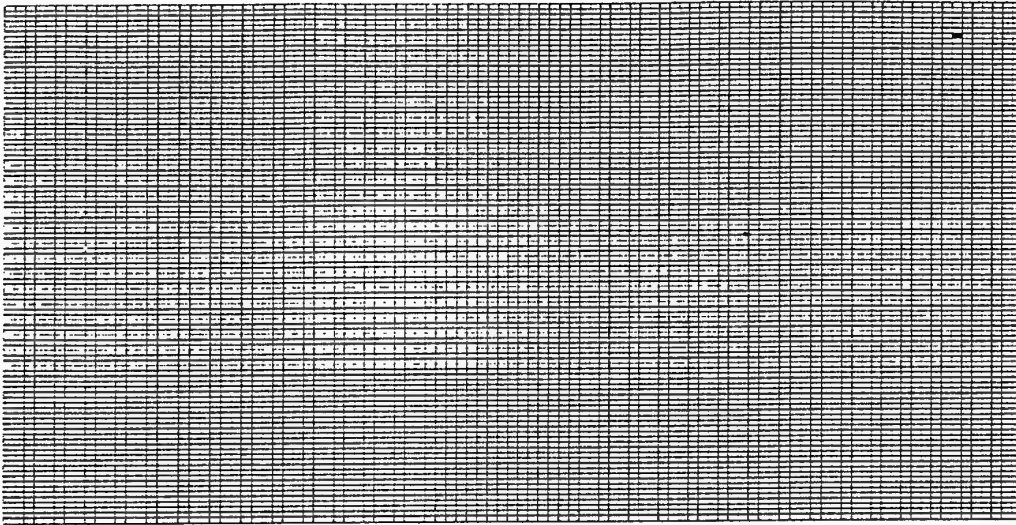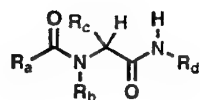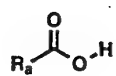
	Ra	C	O	O	H	Rb	N	H	H				Rd	N	C
Ra		1													
C	1		2	1											
O		2	4												
O		1		4	1										
H				1											
Rb							1								
N						1	2	1	1						
H							1								
H							1								
Rd														1	
N													1	2	2
C														2	2

+

R

	Ra	C	O	O	H	Rb	N	H	H				Rd	N	C
Ra															
C			-1				1								
O															
O	-1			-1											2
H			-1												
Rb															
N		1						-1	-1						
H							-1							1	
H							-1								
															1
Rd							1								
N							1								-2
C			2								1			-2	-2

=

B

	Ra	C	O	O	H	Rb	N	H	H	Rc	C	H	O	Rd	N	C
Ra		1														
C	1		2				1									
O		2	4													
O				4												2
H																
Rb							1									
N		1				1	2									
H													1			
H							1									
																1
Rd													1			
N							1						1	2	1	
C			2							1			1			

Fig. 7

a1	a2	a3	a4	a5	a6	a7	a8	a9	b1	b2	b3	b4	b5					d1	d2	d3	d4	d5	d6	d7	ra	...	rb	...	rc	...	rd	...
ε:= ra	a5	a6	a7	a2	a3	a4	a9	a8	rb	b5	b4	b3	b2	rc				rd	d4	d5	d2	d3	d7	d6	a1	...	b1	...			d1	...
ρ:= a1	a2	a3	d6	a5	a6	b3		d5	b1	d3		a4		b4	a9		d7	b2	d1	d2	a8	d4	b5		a7	ra	...	rb	...	rc	...	rd	...
β:= ra	a5	a6	b3	a2	a3	d6	d5		rb	a4		d3	rc	d7	b2		b4	a9	rd	d4	b5	d2	a8	a7		a1	...	b1	...		...	d1	...

ε:= (a1 ra) (a2 a5) (a3 a6) (a4 a7) (a8 a9) (b1 rb) (b2 b5) (b3 b4) (rc) ( ) ( ) ( ) (d1 rd) (d2 d4) (d3 d5) (d6 d7)

ρ:= (b4 d6 a4) (d7 b3 a7) (d3 b2 a8) (d5 b5 a9)

β:= (a1 ra) (a2 a5) (a3 a6) (a4 b3) (a7 d6) (a8 d5) (a9) (b1 rb) (b2) (b4) (b5 d3) (rc) (d7) ( ) (d1 rd) (d2 d4)

A MODEL FOR COMBINATORIAL ORGANIC CHEMISTRY

Sherif El-Basil

Faculty of Pharmacy , Kasr El-Aini street

Cairo 11562 Egypt

ABSTRACT

The set of coset representations , CR's , of a group G , $\{G(/G_1)$, $G(/G_2)$, ... , $G(/G_s)\}$, Where $G_1 = \{I\}$; $G_s = G$; the marks , m_{ij} of subgroup G_j on a given $G(/G_i)$, $1 \le i \le s$, and the subduction of $G(/G_i)$ by G_j, $j \le i$, $G(/G_i) \downarrow G_j$, are essential tools for the enumeration of stereo isomers and their classification according to their subgroup symmetry [Fujita ,S. , Symmetry and Combinatorial Enumeration in Chemistry , Springer - Verlag , Berlin 1991]. In this paper , each $G(/G_i)$ is modelled by a set of coloured equivalent configurations , $\mathcal{H} = \{h_1$, h_2 , ... , $h_r\}$; $r = |G|/|G_i|$, (called homomers) , such that a given homomer , h_k , remains invariant *only* under all $g \in G_i$ where g is an element of symmetry. The resulting homomers generate the corresponding set of marks almost by inspection. The symmetry relations among a set can $\mathcal{H}$ be conveniently stored in a Cayley - like diagram [Chartrand , G, Graphs as Mathematical Models ,Prindle , Weber and Schmidt Incorporated, Boston , MA , 1977 , Chapter 10] , which is a complete digraph on r vertices so that an arc from the vertex v_i to the vertex v_j is coloured with the set S_{ij} of symmetry elements such that $hi \xrightarrow{g_{ij}} h_j$; $g_{ij} \in S_{ij}$.

In addition , each vertex , v_i is associated with a loop which is coloured with a set S_{ij} so that $g_{ij} \in S_{ii}$ stabilizes h_i.

A Cayley- like diagram of a given CR , $\mathcal{D}[G(/G_i)]$ leads to graphical generation of $G(/G_i) \downarrow G_j$ for all values of j and also to all m_{ij}'s.

Several group-theoretical results are rederived and /or became more envisagable through this modelling. The approach is examplified using C_2 , C_3 , D_2 , T and D_3 point groups and is applied to trishomocubane , a molecule which belongs to the D_3 point group.

1.Introduction and Background

Suppose a parent skeleton which belongs to a given point group, G, to be subjected to a particular substitution pattern leading to a number of structures. Let $G_1=C_1(=\{I\})$, G_2, ..., G_s, be the sequence of representative subgroups [1], where $G_s = G$. One may ask : How many derivatives are there which belong to each subgroup ?

In the past few years Fujita has published a number of papers [2-9] in which he developed powerful group - theoretical methods which answer the above and such questions. While conventional treatments consider linear representations and character tables of groups [10], Fujita's approach is based on **coset representations**, CR's , and **table of marks**. [11] Namely, each element of symmetry, $g \in G$ applied to a coset of G_i; $1 \le i \le s$ gives another coset and thus each element of G can be considered as a certain permutation of the cosets , leading to a representation of G in terms of these permutations , called **coset representations** , $G(/G_i)$. Formally :

$$G(/G_i)=\{\pi_g \mid \forall\ g \in G\}\ ;\qquad\qquad ...(1)$$

$$\pi_g = \begin{pmatrix} G_ig_1 & G_ig_2 & ... & G_ig_r \\ G_ig_1g & G_ig_2g & ... & G_ig_rg \end{pmatrix}\qquad ...(2)$$

In eqn (2) , $r = |G| / |G_i|$, $|G|$ = order of G. The **mark** m_{ij} of G_j on $G(/G_i)$, G_j being another subgroup of **G** , is the number of cosets left invariant (fixed) by G_j. The set of vertices of the molecular graph which undergoes substitution is called the **orbit** of substitution. Essential to the treatment of Fujita is to classify the orbit which undergoes substitution according to the CR which governs its substitution [6]. This orbit is then **subduced** by all subgroups of the parent point group in order to obtain the required structural counts which are expressed in the so-called isomer-count matrix[1]. The subduction of a CR , $G(/G_i)$ by G_j is expressed by eqn(3) which is related to eqn(1) by just attaching the subscript j to G in braces , viz. .

$$G(/G_i) \downarrow G_j = \{\ \pi_g \forall \mid g \in G_j\}\qquad\qquad ...(3)$$

The basic tools here (cosets , marks , subduction tables etc...) are indeed rather abstract in nature and may not draw the attention of an organic chemist who benefits most from the **results** of this algebra [12]. Here we propose a graphical modelling of mark and subduction tables of CR's using "Cayley [13] - like colour graphs" * (see below). Due to their diagrammatic nature , graphs are more appealing to chemists and

---

* or simply , Cayley diagrams , for brevity.

in addition the suggested graphical model rederives many of the group theoretical results which are conventionally obtained using purely algebraic methods.

2. Modelling Coset Representations of Groups:

First a molecular graph is drawn which remains fixed under all of the symmetry operations of $\mathcal{D}$. In **FIG** 1 several graphs are drawn which represent models of some of the simpler point groups. The D_2 model is a tapered conformation of ethylene while D_3 is a twisted ethane which is neither staggered nor eclipsed. The **T** point group model is an adamantane molecule graph. Orbits of substitution are represented by the (open) vertices of each graph. When all the vertices are opened the model is said (here) to be "uncoloured" and by definition such a model represents the **regular** coset representation, $G(/G)$, since it remains fixed under all $g \in G$. For the other CR's, $G(/G_i)$; $1 \leq i \leq s$ one is dealing with less symmetry operations, viz., only those which belong to G_i and whence a particular "colouring" of the original model of $G(/G)$ is adopted so that the coloured graph remains fixed **only** under G_i. Arbitrarily the number of black ($\equiv$ closed) vertices is chosen to be a minimum. Further, one must search for all such coloured **equivalent configurations** which reproduce the permutation properties for the cosets of G_i. The chemical term for equivalent configurations is **homomers**. For a given CR, $G(/G_i)$ this set will be denoted as $\mathcal{H}[$ $G(/Gi)$] where :

$$\mathcal{H} [G(/G_i)] \equiv \mathcal{H} = \{h_1, h_2, ..., h_r\} \qquad ...(4)$$

in which h_i is an ith homomer and r is defined by eqn(2). To find $h_1 \in \mathcal{H} [G(/G_i)]$ we apply all g's $\in G_i$ and express the result in cyclic notation [14] such as (ab)(cdef)(...) where the letters in parentheses refer to the labels of vertices of the orbit. Since h_1 remains fixed under G_i it must yield the following "vertex-colour" identities:

$$a=b \; ; c=d=e=f, ... \qquad ...(5)$$

We, then select **just one** such equality which involves the smallest number of vertices and colour them in black. The resulting (coloured) configuration is (arbitrarily) called h_1. The other, (r-1), elements of the set $\mathcal{H}$ are deduced so that they generate the same cyclic structures of the permutations representing $G(/G_i)$. In **Appendix 1**, this is demonstrated for $D_3(/C_2)$. **FIG.** 2 portrays sets of homomers which model C_2, C_3 and D_2 point groups.

FIGs 3 and 4 model all CR's of D_3 and T point groups respectively.

3. Modelling Marks of a Given $G(/G_i)$ Using $\mathcal{H}[G(/G_i)]$

Firstly , one lists below each homomer the subgroups which leave it fixed and then arrange the results in the form of a row vector the entries of which are the representative [1] subgroups of G ordered in a nonascending order , viz. , $|G_1| \leq |G_2| \leq \ ... \leq |G_s|$. In FIGs. 2-4 rows of marks from each CR are shown. The above non ascending sequence is called sequence of subgroups , SSG.

4. Inter-relations Among a Set of Homomers

In order to further our knowledge of the properties of a set $\mathcal{H}[G(/G_i)]$, the inter-relations among the individual members (homomers) can be outlined either graphically or in matrix form.

4.1. Graphical Representation of $\mathcal{H}$:"Cayley [13] - like Colour diagrams"

the set of r homomers $\{h_1, h_2, ..., h_r\} \equiv \mathcal{H}$ which is associated with a given CR , $G(/G_i)$ can be represented by a complete digraph (directed graph) on r vertices with arcs and loops. Each arc say from h_i to h_j is "coloured" , so to speak , with a set of symmetry elements , S_{ij} , such that

$$h_i \xrightarrow{\quad g_{ij} \quad} h_j \, ; \qquad\qquad ...(6)$$
$$g_{ij} \in S_{ij}$$

each vertex , v_i , is associated with a loop which is also coloured with a set of elements , S_{ii} so that a given $g_{ii} \in S_{ii}$ stabilises h_i. The resulting digraphs , $\mathcal{D}[\mathcal{H}(G/G_i)]$ are reminiscent of "Cayley colour graphs" of groups [13] . FIG. 5 shows $\mathcal{D}$ graphs of CR's of some point groups. We shall demonstrate that the $\mathcal{D}$ graphs contain information on:

a) Coset decomposition of G by G_i ;

b) Marks of $G(/G_i)$; and

c) $G(/G_i) \Downarrow G_j$ (c.f. eqn.3) for all G_j of G_i ; where G_j is a subgroup of G_i.

Furthermore , the above information can be extracted almost immediately from these digraphs.

4.2 Matrix Representation of $\mathcal{H}$

If $|\mathcal{H}\,[G(/G_i)]|$ is large (say >4) it may be more convenient to work with $\mathcal{D}$ in its matrix form , $M[G(/G_i)]$. The matrix is defined as an r x r matrix whose diagonal and off-diagonal elements are the sets S_{ii} and S_{ij} respectively (c.f. Section 4.1). **FIG.** 6 shows **M** matrices of several coset representations of some point groups.

4.3 Graphical Representation of eqn(3) : Subduced Representations

$G(/G_i) \downarrow G_j$ is usually expressed as a sum of CR's of G_j , viz. ,

$$G(/G_i) \; \downarrow \; G_j = \alpha_{jk}G_j(/G_k) + \alpha_{jl}G_j(/G_l) +... \qquad ...(7)$$

where G_k , G_l, ... are subgroups of G_j and the α's are non-negative multiplication factors.

The "subduction" set of equivalent configurations described above may be obtained graphically by the following steps:

a) Find the Cayley-like colour graph of $G(/G_i)$, $\mathcal{D}[G(/G_i)]$ as described in section 4.1.

b) an arc (or a loop) in $\mathcal{D}[G(/G_i)]$ is annihilated (i.e. pruned out) unless one of its colour components belongs to G_j , the subducing group.

c) The result of b) is in general , a set of disconnected $\mathcal{D}$'s of the general form

$$\{ \quad \mathcal{D}[G_j(/G_k)] \, . \, \mathcal{D}[G_j(/G_l)]...\} \qquad ...(9)$$

where , in general , each component of the above set may be repeated α times [c.f. eqn.(7)].

d) the result of subduction is obtained by comparing the resulting diagrams with those shown in **FIG**. 5. **FIG** 7 illustrates such a graphical subduction of the CR : $T(/C_3)$ by C_2 and by C_3.

Alternatively the set $G(/G_i) \downarrow G_j$ might be obtained by subducing the matrix $M[G(/G_i)]$ in the following steps:

a) Annihilate from $M[G(/G_i)]$ all $g \notin G_j$ to give $M[G(/G_i) \downarrow G_j]$

b) If $M[G(/G_i) \downarrow G_j]$ is not already in block form , then apply to it the appropriate set of row/column operations [15] to transform it into block form. (The notation H_{ij} (K_{ij}) $\equiv$ interchanging rows (columns) i and j). A general form of the matrix which results in this step is given by:

$$\left(\begin{array}{ccccc} [G_j(/G_k)] & \mathbf{O} & \mathbf{O} & ... & \mathbf{O} \\ \mathbf{O} & [G_j(/G_l)] & \mathbf{O} & ... & \mathbf{O} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & ... & [G_j(/G_m)] \end{array} \right) \qquad ...(10)$$

where G_k , G_l , ... , G_m are subgroups of G_j.

c) The block matrix , (10) , represents the disconnected Cayley-like colour graphs:

$$\{ \mathcal{Y}[G_j(/G_k)] . \mathcal{Y}[G_j(/G_l)] ... \mathcal{Y}[G_j(/G_m)] \} \qquad ...(11)$$

d) Then the matrix given by eqn(10) corresponds to the subduction expression given by eqn.(7).

Example 1

$$T(/C_2) \downarrow D_2 = D_2(/C_2) + D_2(/C_2') + D_2(/C_2'') + ... \qquad ...(12)$$

We observe that $|\mathcal{Y}[T(/C_2)]| = 6$ and whence it is more convenient to work with matrix forms , steps a)-d) are illustrated below:

$$\left(\begin{array}{cccccc} \{I,C_2\} & \{C_2',C_2''\} & \{C_{3(1)},C_{3(3)}\} & \{C_{3(2)},C_{3(4)}\} & \{C'_{3(2)},C'_{3(3)}\} & \{C'_{3(1)},C'_{3(4)}\} \\ \{C_2',C_2''\} & \{I,C_2\} & \{C_{3(2)},C_{3(4)}\} & \{C_{3(1)},C_{3(3)}\} & \{C'_{3(1)},C'_{3(4)}\} & \{C'_{3(2)},C'_{3(3)}\} \\ \{C'_{3(1)},C'_{3(3)}\} & \{C'_{3(2)},C'_{3(4)}\} & \{I,C_2'\} & \{C_2,C_2''\} & \{C_{3(3)},C_{3(4)}\} & \{C_{3(1)},C_{3(2)}\} \\ \{C'_{3(2)},C'_{3(4)}\} & \{C'_{3(1)},C'_{3(3)}\} & \{C_2,C_2''\} & \{I,C_2'\} & \{C_{3(1)},C_{3(2)}\} & \{C_{3(3)},C_{3(4)}\} \\ \{C_{3(2)},C_{3(3)}\} & \{C_{3(1)},C_{3(4)}\} & \{C'_{3(3)},C'_{3(4)}\} & \{C'_{3(1)},C'_{3(2)}\} & \{I,C_2''\} & \{C_2,C_2'\} \\ \{C_{3(1)},C_{3(4)}\} & \{C_{3(2)},C_{3(3)}\} & \{C'_{3(1)},C'_{3(2)}\} & \{C'_{3(3)},C'_{3(4)}\} & \{C_2,C_2'\} & \{I,C_2''\} \end{array} \right)$$

$$M[T(/C_2)]$$

Keep only $g \in D_2$

$$\left(\begin{array}{cccccc} \{I,C_2\} & \{C_2',C_2''\} & O & O & O & O \\ \{C_2',C_2''\} & \{I,C_2\} & O & O & O & O \\ O & O & \{I,C_2'\} & \{C_2,C_2''\} & O & O \\ O & O & \{C_2,C_2''\} & \{I,C_2'\} & O & O \\ O & O & O & O & \{I,C_2''\} & \{C_2,C_2'\} \\ O & O & O & O & \{C_2,C_2'\} & \{I,C_2''\} \end{array} \right)$$

$$M[T(/C_2) \downarrow D_2]$$

$$\sim \begin{pmatrix} [D_2(/C_2)] & O_2 & O_2 \\ O_2 & [D_2(/C'_2)] & O_2 \\ O_2 & O_2 & [D_2(/C''_2)] \\ & M[T(/C_2) \downarrow D_2] & \end{pmatrix}$$

where O_2 is a 2 x 2 null matrix

In this example no row /column operations were necessary , i.e. the subduced matrix was already in block form. The following example represents a more general situation where one must apply a set of row/column operation to obtain the expression of the subduced representation:
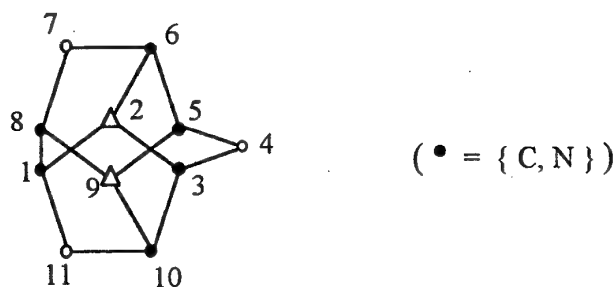
Example 2

$$T(/C_2) \downarrow C_3 = 2 \, C_3(/C_1) \qquad \qquad ...(13)$$

$$[T(/C_2)\,] \xrightarrow[g \in C_3]{\text{Keep only}} \begin{pmatrix} \{I\} & O & \{C_{3(1)}\} & O & O & \{C'_{3(1)}\} \\ O & \{I\} & O & \{C_{3(1)}\} & \{C'_{3(1)}\} & O \\ \{C'_{3(1)}\} & O & \{I\} & O & O & \{C_{3(1)}\} \\ O & \{C'_{3(1)}\} & O & \{I\} & \{C_{3(1)}\} & O \\ O & \{C_{3(1)}\} & O & \{C'_{3(1)}\} & \{I\} & O \\ \{C_{3(1)}\} & O & \{C'_{3(1)}\} & O & O & \{I\} \end{pmatrix}$$

$$\xrightarrow[H_{23},\,H_{36}]{K_{23},\,K_{36}} \begin{pmatrix} [C_3(/C_1)] & O_3 \\ O_3 & [C_3(/C_1)] \end{pmatrix}$$

An application of the model   Find the isomer-count matrix of the following substituted trishomocubane graph.

$$(\bullet = \{ C, N \})$$

$$... (14)$$

Trishomocubane is a caged compound which belongs to the D_3 point-group[16] [c.f. **FIG** 1]. The compound can be envisaged from the fusion of six equivalent cyclopentane rings (or three norbornanes). The parent (unsubstituted graph of this molecule has three C_2 axes of symmetry each passing through a methylene carbon and the center of an opposite edge: One C_2 axis passes through vertex 4 and edge joining vertices 1 and 8, Another C_2 axes passes through vertex 7 and the bond joining vertices 3 and 10. Finally a third C_2 passing through vertex 11 and the bond between

vertices 5 and 6. The molecular also has two C_3 axes, each passes through vertices 1 and 9. Inspection of the graph shown in eqn. (15) reveals three orbits, viz.,

$$\Delta_1 = \{ 4, 7, 11 \} ; \text{ (colored as open circles)} \qquad \dots (15)$$
A subset of bivalent vertices.

$$\Delta_2 = \{ 1, 3, 5, 6, 8, 10 \} ; \text{ (colored as solid circles)} \qquad \dots (16)$$

A subset of trivalent vertices each one is adjacent to two trivalent vertices and one bivalent vertex.

$$\Delta_3 = \{2,9\} ; \text{ (colored as open triangles)} \qquad \dots (17)$$
A subset of trivalent vertices, each vertex of which is surrounded by three trivalent vertices.

FIG. 1 and 3 show how to model coset representations of D_3 point group which lead to a model of the mark table of this group in **FIG.** 8. The conventional (numerical) from is given in **Table** 1.

Table 1

Mark table of D_3 point-group

	C_1	C_2	C_3	D_3
$D_3(/C_1)$	6	0	0	0
$D_3(/C_2)$	3	1	0	0
$D_7(/C_3)$	2	0	2	0
$D_3(/D_3)$	1	1	1	1

Now to define the coset representation which controls the orbit of substitution, Δ_2 in this case, eqn (16), we apply the elements of symmetry of D_3 to the vertices of Δ_2 and count the number of fixed vertices under the effects of all subgroups: the rows of the mark table of the full group (**Table** 1). These operations are shown below:

			C_1	C_2	C_3	D_3
I	$\rightarrow$	(1)(6)(5)(3)(10)(8)	√	√	√	√
C_2	$\rightarrow$	(18)(6 10) (3 5)		√		√
C'_2	$\rightarrow$	(3 10) (15) (68)				√
C''_2	$\rightarrow$	(5 6) (1 10) (38)				√
C_3	$\rightarrow$	(6 1 3) (5 8 10)			√	√
C'_3	$\rightarrow$	(6 3 1) (5 10 8)			√	√
			C_1	C_2	C_3	D_3
			(6	0	0	0)

$$\dots (18)$$

Comparison of the vector generated in eqn. (18) with **mark table**, **Table** 1, indicates that the orbit which controls substitution in Δ_2 is D_3 ($/C_1$) type. This is the CR which must be subduced by the four subgroups of D_3. The resulting USCI's are outlined below together with the corresponding generating functions adopting the weights:

$$w\,(C) = x \quad ; \qquad w\,(N) = y \qquad \dots (19)$$

$$\downarrow C_1 \;\rightarrow\; S_1^6 \;\rightarrow\; (x + y)^6 \qquad \dots (20)$$

$$\downarrow C_2 \;\rightarrow\; S_2^3 \;\rightarrow\; (x^2 + y^2)^3 \qquad \dots (21)$$

$$\downarrow C_3 \;\rightarrow\; S_3^2 \;\rightarrow\; (x^3 + y^3)^2 \qquad \dots (22)$$

$$\downarrow D_3 \;\rightarrow\; S_6 \;\rightarrow\; (x^6 + y^6) \qquad \dots (23)$$

An illustration of the subduction $D_3(/C_1) \downarrow C_2 = 3\; C_2(/C_1)$

The Cayley graph of the CR $D_3(/C_1)$ has six vertices which represent its six homomer models. This graph is too large to construct and then apply the pruning technique to expand the required subduction . In this and similar situations the matrix

representation of CR is a more convenient method. The homomers of $D_3(/C_1)$ shown in **FIG.3** transform to one another according to the following matrix:

$$M[D_3(/C_1)] = \begin{array}{c} \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \end{array} \begin{array}{cccccc} h_1 & h_2 & h_3 & h_4 & h_5 & h_6 \\ \left(\{ I \} \ \{C_2\} \ \{C_2'\} \ \{C_2''\} \ \{C_3\} \ \{C_3'\} \right. \\ \{C_2\} \ \{ I \} \ \{C_3'\} \ \{C_3\} \ \{C_2''\} \ \{C_2'\} \\ \{C_2'\} \ \{C_3\} \ \{ I \} \ \{C_3'\} \ \{C_2\} \ \{C_2''\} \\ \{C_2''\} \ \{C_3'\} \ \{C_3\} \ \{ I \} \ \{C_2'\} \ \{C_2\} \\ \{C_3'\} \ \{C_2''\} \ \{C_2\} \ \{C_2'\} \ \{ I \} \ \{C_3\} \\ \left. \{C_3\} \ \{C_2'\} \ \{C_2''\} \ \{C_2\} \ \{C_3'\} \ \{ I \} \right) \end{array} \quad \ldots (24)$$

To obtain the required expression for subduction (with C_2) we keep only those elements of $C_2(= \{I, C_2\})$ in the above matrix and carry out the appropriate row/column operations as shown below:

$$\begin{array}{c} \\ h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \end{array} \begin{array}{cccccc} h_1 & h_2 & h_3 & h_4 & h_5 & h_6 \\ \left(I & C_2 & 0 & 0 & 0 & 0 \right. \\ C_2 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & C_2 & 0 \\ 0 & 0 & 0 & I & 0 & C_2 \\ 0 & 0 & C_2 & 0 & I & 0 \\ \left. 0 & 0 & 0 & C_2 & 0 & I \right) \end{array} \quad = M[D_3(/C_1) \downarrow C_2] \quad \ldots (25)$$

$$\downarrow \begin{array}{l} 1.K_{45} \\ 2.H_{35} \end{array}$$

$$\begin{pmatrix} I & C_2 & 0 & 0 & 0 & 0 \\ C_2 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & I & C_2 & 0 & 0 \\ 0 & 0 & C_2 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & I & C_2 \\ 0 & 0 & 0 & 0 & C_2 & I \end{pmatrix} \quad \equiv \quad \begin{pmatrix} C_2(/C_1) & 0_2 & 0_2 \\ 0_2 & C_2(/C_1) & 0_2 \\ 0_2 & 0_2 & C_2(/C_1) \end{pmatrix} \quad \ldots (26)$$

where $K_{ij} \equiv$ interchange column i and column j while $H_{ij} \equiv$ interchange row i and row j and 0_2 is a 2 ×2 null matrix. Eqn.(26), then corresponds to three $C_2(/C_1)$'s which is

what one obtains using conventional coset algebra which requires both permutation representations **and** mark tables of the subducing groups.

The resulting polynomials , eqns.(19)-(23),when expanded generate the following FP matrix:

Table 2

    Fixed-point matrix of the trishomocubane problem.

	C_1	C_2	C_3	D_3
x^6	1	1	1	1
y^6	1	1	1	1
xy^5	6	0	0	0
x^5y	6	0	0	0
x^2y^4	15	3	0	0
x^4y^2	15	3	0	0
x^3y^3	20	0	2	0

When the mark table is applied into **Table 2** we obtain the desired isomer-count matrix, shown below:

Table 3

    Isomer-count matrix of the trishomocubane graph.

	C_1	C_2	C_3	D_3
x^6				$*^1$
y^6				$*^2$
xy^5	$*^3$			
x^5y	$*^4$			
x^2y^4	$*^5$	$*^6 *^7 *^8$		
x^4y^2	$*^9$	$*^{10} *^{11} *^{12}$		
x^3y^3	$*^{13} *^{14} *^{15}$		$*^{16}$	

Where each star corresponds to a row from the mark table. The labelings of the stars correspond to the heterocycles derived from this caged molecule. These are drawn in **FIG** 9.

5. Modeling Algebraic Properties of Mark and Group - Subduction -Tables

5.1 Properties of Marks

5.1.1 The row of marks of $G(/G_1)$ has the general form

$$(|G| \quad O \quad O \quad ... \quad O) \qquad ...(27)$$

where the number of O's $=|SSG|-1$. This is because every loop in $\mathcal{Y}[G(/G_1)]$ is coloured with just one component , viz. , C_1 and recalling that $|G(/G_1)| = |G| = |G[G(/G_1)]|$ the general form of row of marks given by eqn.(24) results.

5.1.2 The row of marks of $G(/G)$ has the general form :

$$(1 \quad 1 \quad ... \quad 1) \qquad ...(28)$$

Where the number of 1's $= |SSG|$. This property which one observes at the bottoms of mark Tables results from the fact that $\mathcal{Y}[G(/G)]$ has the general form of a single vertex whose loop is coloured with every subgroup of G. I.e. , takes the general form:

$$\{C_1, G_2,...,G\}$$

$$...(29)$$

5.1.3 The row of marks of a given $G(/G_i)$ has only two values , viz. ,

$$m_{ij} = \left\{ \begin{array}{l} O \\ \\ r \end{array} \right. \qquad ...(30)$$

iff all the loops of $\mathcal{Y}[G(/G_i)]$ are identically coloured with the same subgroups. Examples are shown in **FIG. 2** for $D_2(/C_2)$; $D_2(/C'_2)$ and $D_2(/C''_2)$.

5.2 Properties of Subduced Representations

5.2.1 There are two dominant characters of the general sum given by eqn. (7) namely :

a)The number of homomers which represents (models) a given coset representation ,

$G(/G_i)$, = the number of homomers which models a given subduced representation of $G(/G_i)$ by one of its subgroups' i.e.:

$$|\mathcal{H}[G(/G_i)]| = |\mathcal{H}[G(/G_i) \downarrow G_j]| \qquad \ldots(31)$$

b) The two sets of homomers of eqn.(28) have identical transformation properties under all $g \in G_j$ i.e. under all symmetry operations of the subducing groups.

These two properties may be modeled by considering for example:

$T(/C_3) \downarrow C_3 = C_3(/C_1) + C_3(/C_3)$; $C_3 = \{I, C_3, C_3'\}$

$$= C_3(/C_3) + C_3(/C_1) \qquad \ldots(32)$$

Property a) is also understood using graphical subduction of the Cayley diagram since the total number of vertices in the fragmented graph is preserved.

5.2.2 Subduction of the identity representation of a group leads to the identity representation of the subducing group , i.e. ,

$$G(/G) \downarrow G_j = G_j(/G_j) \qquad \ldots(33)$$

This result is understood from the general form of the Cayley diagram of $G(/G)$, being a single vertex the loop of which is coloured with all the subgroups of G. Then subduction by G_1 leaves only the colour component G_1 to give $G_1(/G_1)$, while in general subduction by G_i leads to a vertex whose loop is coloured by all the subgroups of G_i which corresponds to $G_i(/G_i)$ and so on. This property is modelled below for the identity representation of the **T** point group.

$$\{C_1\} \qquad \qquad \qquad \{C_1, C_3\}$$

$$C_1(/C_1) \xleftarrow{\;\downarrow C_1\;} \{C_1 C_2, C_3, D_2, T\} \xrightarrow{\;\downarrow C_3\;} C_3(/C_3)$$

$$T(/T)$$

$$\{C_1, C_2\} \qquad \qquad \{C_1, C_3, C_2', C_2'', D_2\}$$

$$C_2(/C_2) \xleftarrow{\;\downarrow C_2\;} \qquad \xrightarrow{\;\downarrow D_2\;}$$

$$D_2(/D_2) \qquad \qquad \dots(34)$$
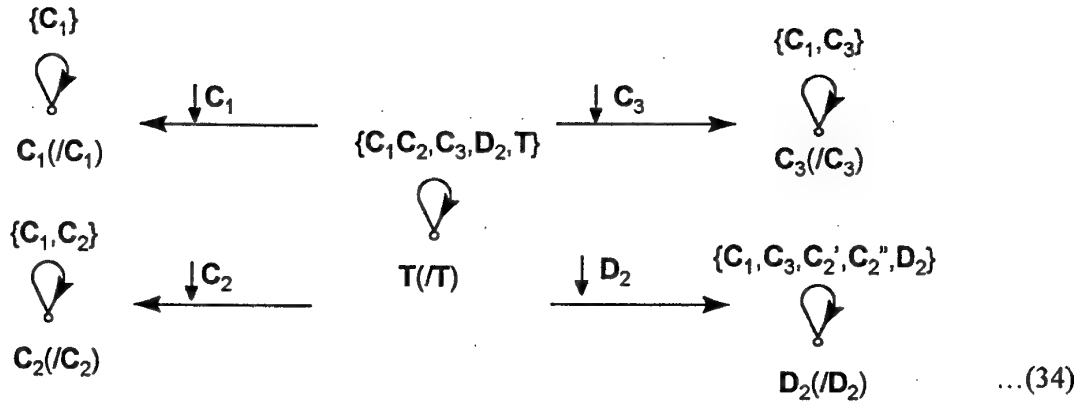
5.2.3 $\qquad\qquad G(/G_1) \downarrow G_j = rG_j(/G_1)$ $\qquad\qquad\dots(35)$

Here the Cayley graph of $G(/G_1)$ possesses loops which are coloured by C_1 only and because C_1 is a common subgroup of all subgroups , the resulting subduced Cayley graphs will also have loops which are coloured by C_1 only and therefore they will be all identically regular representations of the subducing group , i.e. $G_j(/G_1)$. And because $r = |G| / |G_j|$, there will be r such $\mathcal{B}[G_j(/G_1)]$.

5.2.4 $\qquad\qquad G(/G_1) \downarrow G_1 = |G|G_1(/G_1)$ $\qquad\qquad\dots(36)$

Eqn(33) is understood from the fact that the Cayley graph $\mathcal{B}[G(/G_1)]$ contains $|G|$ vertices , the loops of each of which is coloured with G_1 while the arcs with $g \notin G_1$. Then subduction by G_1 fragments it into $|G|$ vertices each of which is nothing else but $G_1(/G_1)$. Eqns (35) and (36) are modelled below:
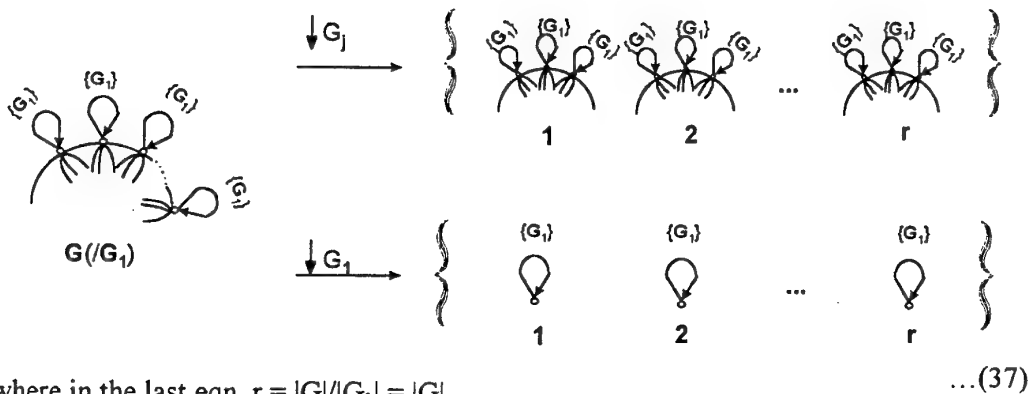

where in the last eqn. $r = |G|/|G_1| = |G|$. $\qquad\qquad\dots(37)$

5.2.5 $$G(/G_i) \downarrow G_i = r\, G_i(/G_i) \qquad \ldots(38)$$

Iff all $h_i \in \mathcal{H}[G(/G_i)]$ remain invariant under G_i.

In this particular case the loops of the vertices of $\mathcal{h}[G(/G_i)]$ are coloured with all the subgroups of G_i (because by assumption all r homomers of $\mathcal{H}[G(/G_i)]$ remain invariant under all $g \in G_i$) and whence none of the arcs are coloured with any $g \in G_i$. Then subduction by G_i generates r lonely vertices whose loops are coloured with all subgroups of G_i, i.e. by $\{SSG_i\}$ which is nothing else but r copies of the regular representation of G_i. This case is modelled below:
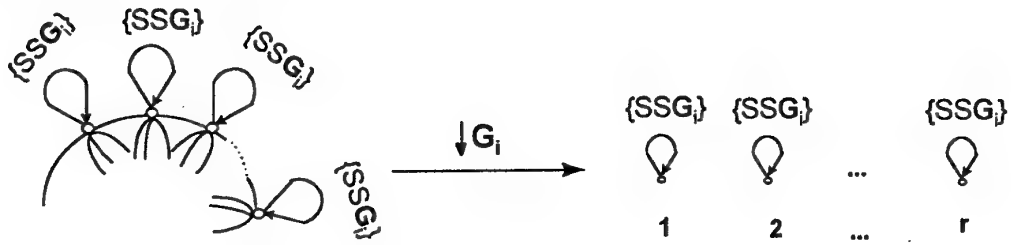
$$\ldots(39)$$

This property is exemplified in **FIG.** 10

*   *   *

6.Discussion and Conclusions

Scheme 1 outlines the main feature of this work:

Point Group **G**

↓

Sequence of representative subgroups :{**G**$_1$, **G**$_2$, ... , **G**$_s$}; **G**$_s$=**G**

↓

Set of Cosets : {**G**(/**G**$_1$) , **G**(/**G**$_2$) , ... , **G**(/**G**)}

↓

Molecular graph which remains invariant under all g ∈ **G**

~ Model of the regular representation , **G**(/**G**)

↓

Colouring of the model of **G**(/**G**) so that it remains fixed

only under g ∈ **G**$_i$; 1 ≤ i < s

↓

A set of equivalent (coloured) configurations (homomers)

for each **G**(/**G**$_i$) ~ $\mathcal{H}$[**G**(/**G**$_i$)].

↓

Mark table ← Cayley colour diagram for each **G**(/**G**$_i$) → Subduction table

Scheme 1

Abstract description of a physical phenomenon remains a most precise and exact description while a given model is always going to be approximate. C.f. , the three (popular) physical chemistry models of an ideal gas , ideal electrolyte and ideal solution. However modelling usually carries both educational as well as theoretical endeavours. In the present work several group theoretical properties are rederived or became more easily envisagable through modelling. Namely , eqn. 3 which defines subductions of a given CR and the concept of a mark : **FIG**.7 is a pictorial illustration of eqn. 3. **FIG. 8** portrays a certainly more appealing form of the mark table (of **D**$_3$ as an example ; c.f. **Table** 1). We believe that the preparation of Cayley diagrams of CR's of the point groups of chemical interest is a worthy task and that such an appendix in a text which deals with enumeration of chemical structures is at least as important as appendices which contain mark tables , subduction tables and their "predecessors" : Character tables !

In conclusion , the model presented here translates the three basic (and abstract) alphabets used in the "enumeration journey" of Fujita, viz., coset , representation , mark and subduction of a group into the language of graphs. The latter shall always remain the more appealing for chemists, Indeed the present work extends an invitation to organic chemists who would like to see organic chemistry from a theoretical (computational) vent but who are also repelled by the (sometimes) offensive algebra in their way. It may be convenient to end this paper by a parody of famous Greek myth (Rex Warner , "Men and Gods" , Kenkyusha , Tokyo 1959)[17].

"An organic chemist demanded to know the riddle and the sphinx said : "What is it that controls elements in a group, controls atoms in a compound and finally isomers in organic chemistry ?" "Is it a coset representation or a mark ? replied the organic chemist. The sphinx found that her riddle was at last answered and died as was fated. The organic chemist received his award and he was made King of the heaven!

Appendix 1

One - to one Correspondence between cosets of a given representation of a group and the corresponding set of homomers illustrated for $D_3(/C_2)$. The permutations resulting from the effect of symmetry elements are outlined below:

$D_3(/C_2) = C_2C_2 + C_2C_3 + C_2C'_3$

	1	2	3		C_1	C_2	C_3	C_4
I	1	2	3	(1)(2)(3)	√	√	√	√
C_2	1	3	2	(1)(23)		√		√
C_2'	3	2	1	(13)(2)				√
C_2''	2	1	3	(12)(3)				√
C_3	2	3	1	(123)			√	√
C_3'	3	1	2	(132)			√	√

$\rightarrow$ row of Marks : (3 1 0 0)

The following correspondences are observed : (c.f. **FIG**.3)

$$C_2C_2 = \{I,C_2\} \sim h_1 \qquad \qquad ...(A-1)$$
$$C_2C_3 = \{C_3,C_2''\} \sim h_2 \qquad \qquad ...(A-2)$$
$$C_2C'_3 = \{C_3',C_2'\} \sim h_3 \qquad \qquad ...(A-3)$$

The elements of symmetry of D_3 generate identical cyclic structures when operate on the homomers h_1 - h_3

	h_1	h_2	h_3	
I	h_1	h_2	h_3	(1)(2)(3)
C_2	h_1	h_3	h_2	(1)(23)
C'_2	h_3	h_2	h_1	(2)(13)
C''_2	h_2	h_1	h_3	(3)(12)
C_3	h_2	h_3	h_1	(123)
C'_3	h_3	h_1	h_2	(132)

Refrences

1. Because conjugate subgroups are considered to be equivalent , only a representative form every set of such subgroups is selected . See e.g., Fujita ,S. , Symmetry and Combinatorial Enumeration in Chemistry ,Springer-Verlag , Berlin **1991** , p.13.
2. Fujita , S., Point groups based on methane and adamantane (T_d) skeletons, J.Chem.Educ., **1986**, 63 , 744-746.
3. Fujita , S., Subduction of coset representations. An application to enumeration of chemical structures, Theor.Chim.Acta, **1989** , 76 , 247-268.
4. Fujita, S., Enumeration of chemical structures by subduction of coset representations. Correlation of unit subduced cycle indices to Pólya's cycle indices , J.Math.Chem., **1990** ,5, 99-120.
5. Fujita , S., Application of coset representations to the construction of symmetry adapted functions , Theor.Chim.Acta , **1990** , 78 , 45-63.
6. Fujita , S., Systematic classification of molecular symmetry by subduction of coset representation , Bull.Chem.Soc.Jpn., **1990** , 63 , 315-327.
7. Fujita , S., Subductive and inductive derivation for designing molecules of high symmetry , J.Chem.Inf.Comput.Sci., **1991** , 31, 540-546.
8. Fujita , S., Unit subduced cycle indices for combinatorial enumeration , J.Graph.Theory, **1994** ,18, 349-371.
9. Fujita , S., promolecules with subsymmetry of $D\infty h$. Combinatroial enumeration and sterochemical properties, J.Chem.Inf.Comput.Sci, **1992** , 32 , 354-363.
10. See e.g. , Tsukerblat , B.S. , Group Theory in Chemistry and Spectroscopy , Academic pree , London , 1994 ; R.L.Flurry , Jr., Symmetry Groups:Theory and Chemical Applications , Prentice-Hall , Inc,, New Jersey , 1980 ;G.Davidson , Group Theory for Chemists , Macmillan , London , 1991.
11. Mark tables , tables of subductions of groups and other useful tables are found in : Fujita , S., Symmetry and Combinatorial Enumeration in Chemistry , Springer-Verlag , Berlin **1991** , pp.321-358.
12. Mead , C.A., Comment on symmetry and combinatorial enumeration in chemistry by Shinsaku Fujita , J.Am.Chem.Soc., **1992** , 114, 4019-4020.
13. Chartrand , G., Graphs as mathematical models, Prindle , Weber & Schmidt , Incorporated , Boston , Ma , **1977** , Chapter 10 , p.232.
14. Cyclic notations are explained in :D.I.A. Cohen , Basic Techniques of Combinatorial Theory , John Wiley & Sons , New York **1978** , Chapter 6.
15. I.N.Herstein and D.J. Winter , A primer on Linear Algebra , Macmillan Publishing Company , New York 1988 , pp.107-119.
16. Underwood G.R. and Ramamoorthy B., Chemical Studies of Caged Compounds , The Synthesis of Pentacyclo [$6,3,0,0^{2,6},0^{3,10},0^{5,9}$] undecane , Tetrahedron lett **47**, pp.4125-4127 (1970).
17. Fujita , S., Private communication 1996.

Figure Legends

Fig.1) Model graphs of five point groups. Some elements of symmetry are shown. The (open) vertices represent orbits of substitution. The indicated graphs model the identity coset representations , $G(/G)$ of point groups.

Fig.2) Sets of homomers (equivalent configurations) which model coset representations of C_2 , C_3 and D_2 point groups. The indicated row vectors are mark rowa corresponding to each coset representation.

Fig.3) Modelling coset representations of D_3 point group. The coloured graphs in braces are homomers which generate the indicated mark rows.

Fig.4) Coset representation of T point group modelled by the appropriate set of homomers along the corresponding mark rows. In all cases the set of homomers which models $G(/G_i)$ remains fixed *only* under a symmetry elements of G_i.

Fig.5) Cayley colour diagrams $\mathcal{b}[\mathcal{H}(G/G_i)]$ which represent coset representations of several point groups. Both arcs and loops are coloured by appropriate (sub-) groups. The number of vertices in each graph = the number of homomers which model the corresponding coset representation. Mark rows are indicated. Observe that when $G_i = C_1$, $\mathcal{b}$ is a single vertex the loop of which is coloured with all subgroups of G. When $G_i = G$ the size of $\mathcal{b} = |G|$.

Fig.6) Matrix representations of coset representations of several point-groups ; c.f. section 4.1.

Fig.7) Graphical modelling of eqn.(3) illustrated for $T(/C_3)$ ↓ C_3 ; $T(/C_3)$ ↓ C_2 through the use of the Cayley colour diagram of $T(/C_3)$.

Fig.8) A "graphical form" of the mark table of D_3. The mark corresponds to a given subgroup is the number of homomers drawn under this subgroup , where ϕ is an empty set. Observe that $h_i \in \mathcal{H}[G(/G_i)]$ remains invariant under any $g \in G_i$; g being a symmetry element. The graphical form of mark table makes the properties of marks more visible.

Fig.9) Heterocyclic derivatives derived from trishomocubane which corresponds to **Table 3** (the isomer-count matrix).
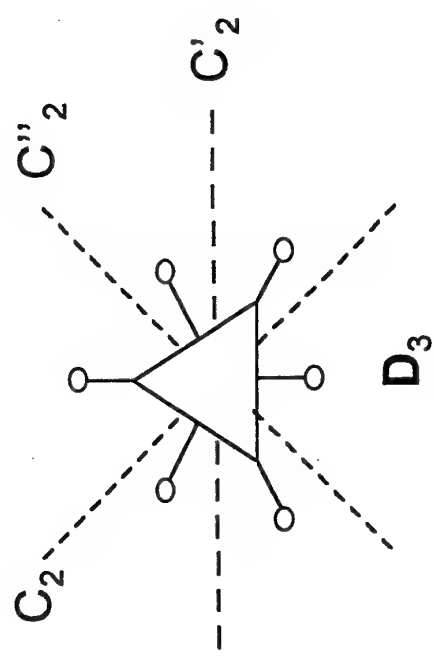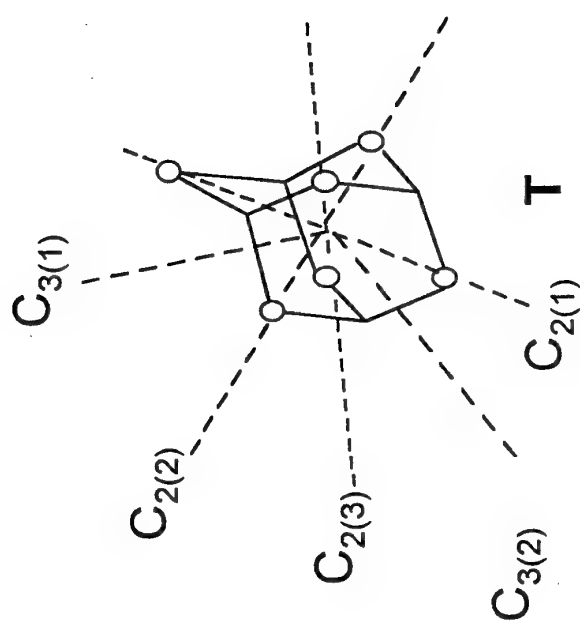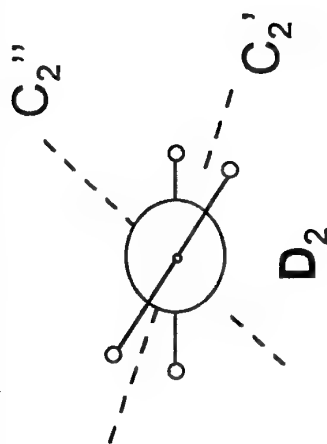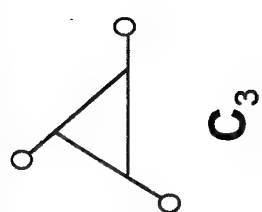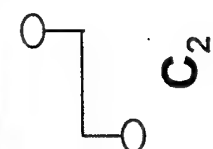
FIG 1

FIG 2 (cont)
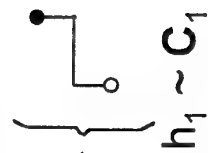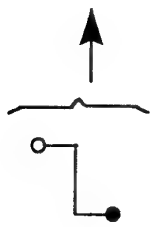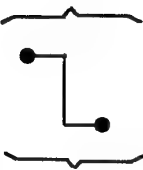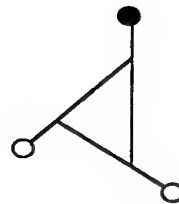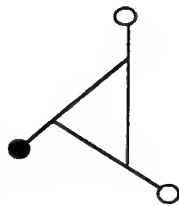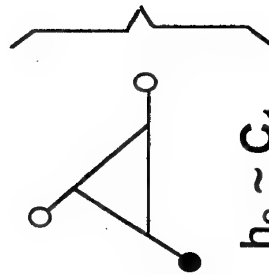
$D_2 = \{1, C_{2(1)}, C_{2(2)}, C_{2(3)}\}$

$D_2(/C_1)$
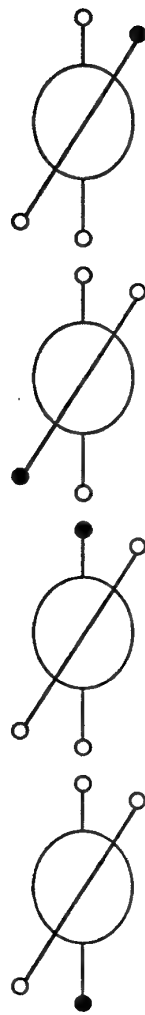
$h_1 \sim C_1$ $h_2 \sim C_1$ $h_3 \sim C_1$ $h_4 \sim C_1$
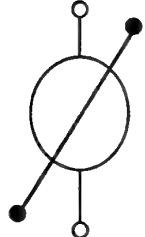
C_1 C_2 C'_2 C''_2

$(4$ 0 0 0 $0)$

$D_2(/C_2)$

$h_1 \sim C_1$ $h_2 \sim C_1$

C_2 C_2

$(2 \quad 2 \quad 0 \quad 0)$

$D_2(/C'_2)$

$h_1 \sim C_1$ $h_2 \sim C_1$

C'_2 C'_2

$(2 \quad 0 \quad 2 \quad 0)$

$D_2(/C''_2)$
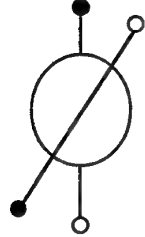
$h_1 \sim C_1$ $h_2 \sim C_1$
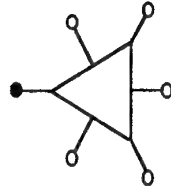
C''_2 C''_2

$(2 \quad 0 \quad 0 \quad 2 \quad 0)$

$D_3 = \{I, C_2, C'_2, C''_2, C_3, C_3{}^2\}$
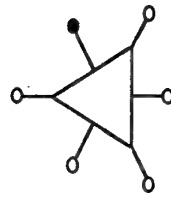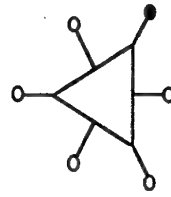
$D_3(/C_1)$
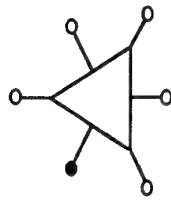
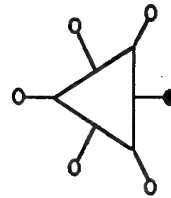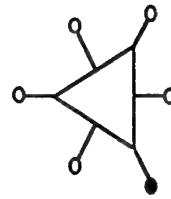$h_1 \sim C_1$ $h_2 \sim C_1$ $h_3 \sim C_1$ $h_4 \sim C_1$ $h_5 \sim C_1$ $h_6 \sim C_1$
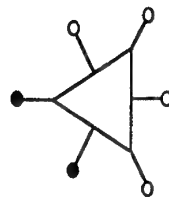
C_1	C_2	C_3	D_3
(6	0	0	0)

$D_3(/C_2)$

$h_1 \sim C_1$ $h_2 \sim C_1$ $h_3 \sim C_1$
C_2

C_1	C_2	C_3	D_3
(3	1	0	0)

$D_3(/C_3)$

$h_1 \sim C_1$ $h_2 \sim C_1$
C_3 C_3

C_1	C_2	C_3	D_3
(2	0	2	0)

FIG. 4
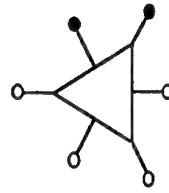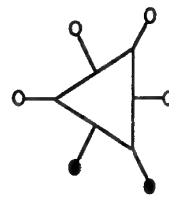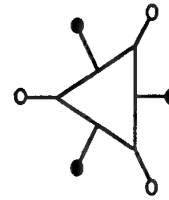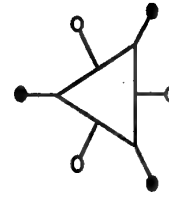
T(/C₂)
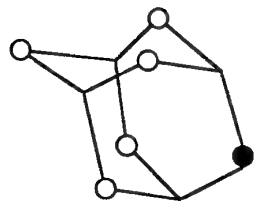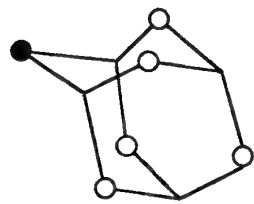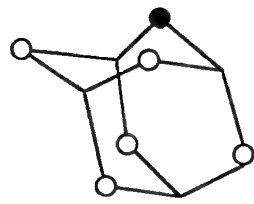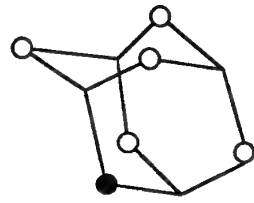
h₁ ~ C₁
   C₂

h₂ ~ C₁
   C₂

h₃ ~ C₁

h₄ ~ C₁
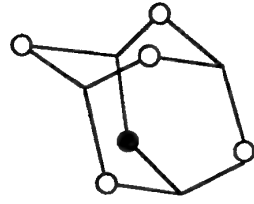
h₅ ~ C₁
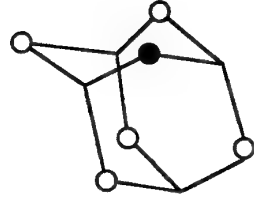
h₆ ~ C₁

C₁   C₂   C₃   D₂   T
(6   2   0   0   0)

T(/C₂) [An alternative model]

h₁   h₂   h₃   h₄   h₅   h₆

$T(/C_3)$

$\sim C_1, C_3$

h_1

$\sim C_1$

h_2

$\sim C_1$

h_3

$\sim C_1$

h_4

$(\;4 \quad 0 \quad 1 \quad 0 \quad 0\;)$

FIG.4
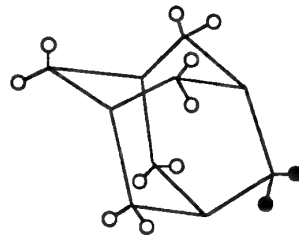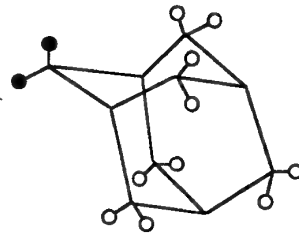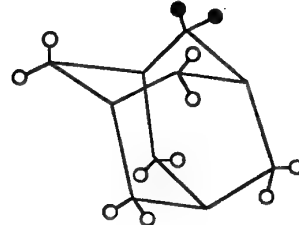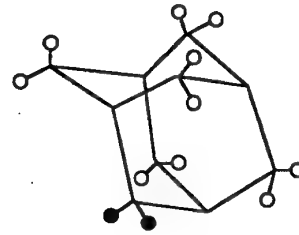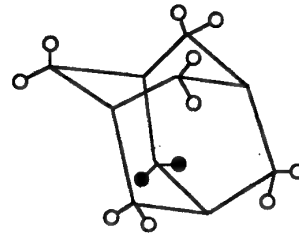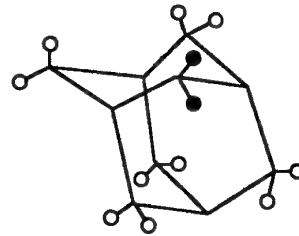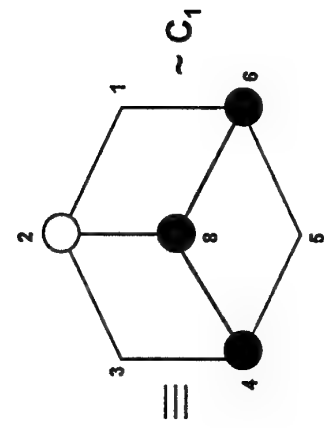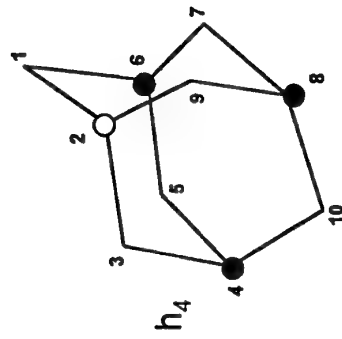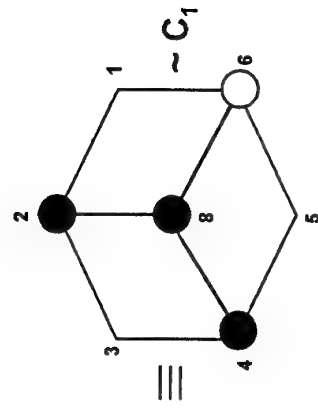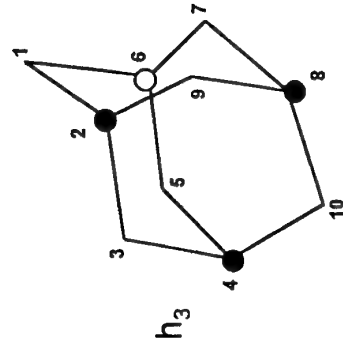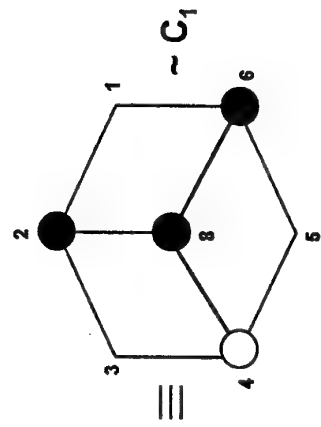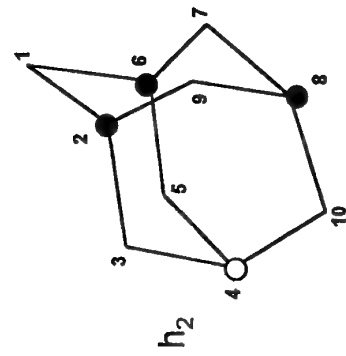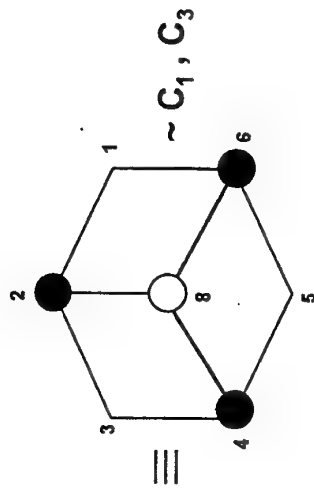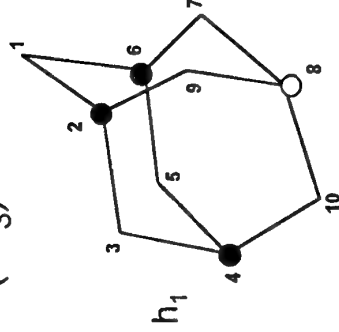
$T(/D_2)$
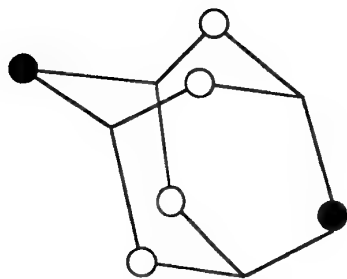
$h_1 \sim C_1$
C_2
D_2
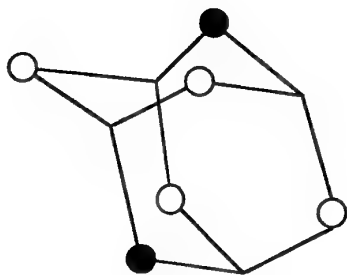
$h_2 \sim C_1$
C_2
D_2

$h_3 \sim C_1$
C_2
D_2

(3   3   0)

    3   0   3

       0

FIG 9
(cont.)

$\{C_1, C_2\}$

$C_2(/C_2)$
(1 1)

$\{C_1, C_3\}$

$C_3(/C_3)$
(1 1)

$\{C_1\}$

$\{C_2\}$ $\{C_2\}$

$\{C_1\}$

$C_2(/C_1)$
(2 0)

$\{C_1\}$

$C_1(/C_1)$
(1)

$\{C_1\}$

$\{C_3\}$ $\{C'_3\}$

$\{C'_3\}$ $\{C_3\}$

$\{C_3\}$ $\{C'_3\}$

$\{C_1\}$

$C_3(/C_1)$
(3   0)

$\{C_1\}$

$\{C_1, C_2\}$ $\{C_1, C_2\}$

$\{C'_2, C''_2\}$ $\{C'_2, C''_2\}$

$D_2(/C_2)$

(2 2 0 0 0)

$\{C_1, C'_2\}$ $\{C_1, C'_2\}$

$\{C'_2, C''_2\}$ $\{C'_2, C''_2\}$

$D_2(/C'_2)$

(2 0 2 0 0)

$\{C_1, C''_2\}$ $\{C_1, C''_2\}$

$\{C_2, C'_2\}$ $\{C_2, C'_2\}$

$D_2(/C''_2)$

(2 0 0 2 0)

$\{C_1\}$ $\{C_1\}$

$\{C''_2\}$ $\{C'_2\}$ $\{C'_2\}$

$\{C''_2\}$ $\{C_2\}$ $\{C_2\}$ $\{C''_2\}$

$\{C'_2\}$ $\{C_2\}$ $\{C_2\}$ $\{C''_2\}$

$\{C'_2\}$ $\{C'_2\}$

$\{C_1\}$ $\{C_1\}$

$D_2(/C_1)$

(4 0 0 0 0)
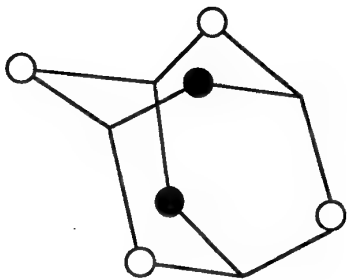
FIG 5
(cont)
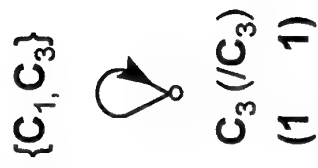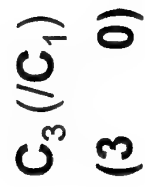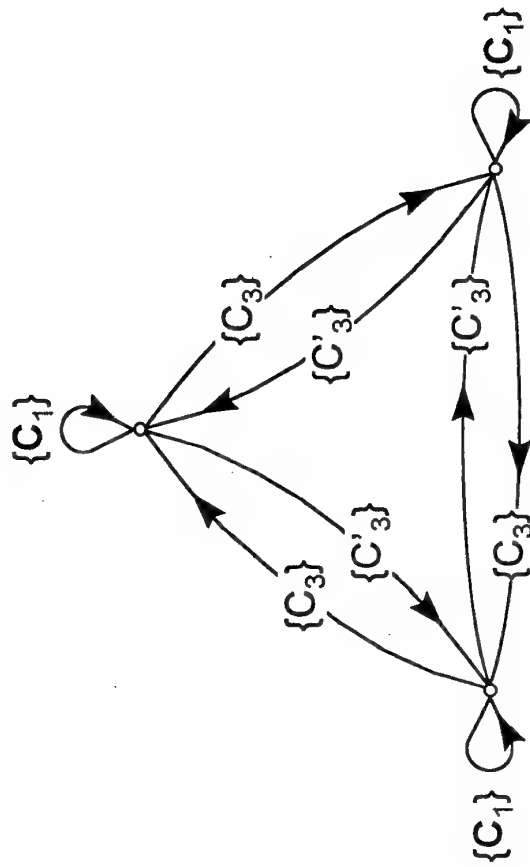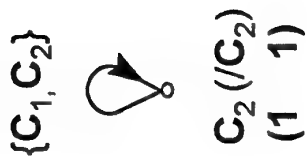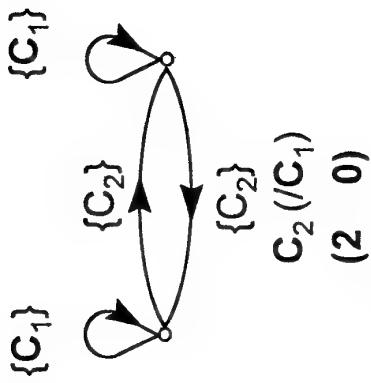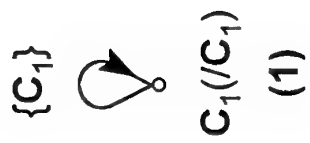
$$\begin{pmatrix} \{C_1\} & \{C_2\} \\ \{C_2\} & \{C_1\} \end{pmatrix}$$

$$\mathbf{M}[C_2\ (/C_1)]$$
$$(2 \quad 0)$$

$$[\{C_1\ C_2\}]$$
$$\mathbf{M}\ [C_2\ (/C_2)]$$
$$(1 \quad 1)$$

$$\begin{pmatrix} \{C_1\} & \{\sigma_v\} \\ \{\sigma_v\} & \{C_1\} \end{pmatrix}$$

$$\mathbf{M}\ [C_s\ (/C_1)]$$
$$(2 \quad 0)$$

$$[\{C_1,\ C_s\}]$$
$$\mathbf{M}\ [C_s\ (/C_s)]$$
$$(1 \quad 1)$$

$$\begin{pmatrix} \{C_1\} & \{C_2\} & \{S_4\} & \{S_4'\} \\ \{C_2\} & \{C_1\} & \{S_4'\} & \{S_4\} \\ \{S_4'\} & \{S_4\} & \{C_1\} & \{C_2\} \\ \{S_4\} & \{S_4'\} & \{C_2\} & \{C_1\} \end{pmatrix}$$

$$\mathbf{M}\ [S_4\ (/C_1)]$$
$$(4 \quad 0 \quad 0)$$

$$\begin{pmatrix} \{C_1, C_2\} & \{S_4, S_4'\} \\ \{S_4, S_4'\} & \{C_1, C_2\} \end{pmatrix}$$

$$\mathbf{M}\ [S_4\ (/C_1)]$$
$$(2 \quad 2 \quad 0)$$

$$(\{\ C_1,\ C_2,\ S_4\})$$
$$\mathbf{M}\ [S_4\ (/S_4)]$$
$$(1 \quad 1 \quad 1)$$

$$\begin{pmatrix} \{C_1\} & \{C_2\} & \{\sigma_v\} & \{\sigma_h\} \\ \{C_2\} & \{C_1\} & \{\sigma_h\} & \{\sigma_v\} \\ \{\sigma_v\} & \{\sigma_h\} & \{C_1\} & \{C_2\} \\ \{\sigma_h\} & \{\sigma_v\} & \{C_2\} & \{C_1\} \end{pmatrix}$$

$$\mathbf{M}\ [C_{2v}\ (/C_1)]$$
$$(4 \quad 0 \quad 0 \quad 0 \quad 0)$$

FIG 6

$$\begin{pmatrix} \{C_1, C_2\} & \{\sigma_v, \sigma_h\} \\ \{\sigma_v, \sigma_h\} & \{C_1, C_2\} \end{pmatrix}$$

$$\mathbf{M}\ [C_{2v}\ (/C_2)]$$
$$(2 \quad 2 \quad 0 \quad 0 \quad 0)$$

$$\begin{pmatrix} \{C_1,C_s\} & \{\sigma_h,C_2\} \\ \{\sigma_h,C_2\} & \{C_1,C_s\} \end{pmatrix}$$

$$\mathbf{M}\,[C_{2v}\,(/C_s)]$$
$$(2\ \ 0\ \ 2\ \ 0\ \ 0)$$

$$\begin{pmatrix} \{C_1,C_s'\} & \{\sigma_v,C_2\} \\ \{\sigma_v,C_2\} & \{C_1,C_s'\} \end{pmatrix}$$

$$\mathbf{M}\,[C_{2v}\,(/C_s')]$$
$$(2\ \ 0\ \ 0\ \ 2\ \ 0)$$

$$\begin{pmatrix} \{C_1\} & \{C_2\} & \{C_2'\} & \{C_2''\} \\ \{C_2\} & \{C_1\} & \{C_2''\} & \{C_2'\} \\ \{C_2'\} & \{C_2''\} & \{C_1\} & \{C_2\} \\ \{C_2''\} & \{C_2'\} & \{C_2\} & \{C_1\} \end{pmatrix}$$

$$\mathbf{M}\,[C_{2v}\,(/C_s')]$$
$$(4\ \ 0\ \ 0\ \ 0\ \ 0)$$

$$\begin{pmatrix} \{C_1,C_2\} & \{C_2',C_2''\} \\ \{C_2',C_2''\} & \{C_1,C_2\} \end{pmatrix}$$

$$\mathbf{M}\,[D_2\,(/C_2)]$$
$$(2\ \ 2\ \ 0\ \ 0\ \ 0)$$

FIG 6
(cont,.)

$$\begin{pmatrix} \{C_1,C_2'\} & \{C_2,C_2''\} \\ \{C_2,C_2''\} & \{C_1,C_2'\} \end{pmatrix}$$

$$\mathbf{M}\,[D_2\,(/C_2'')]$$
$$(2\ \ 0\ \ 0\ \ 2\ \ 0)$$
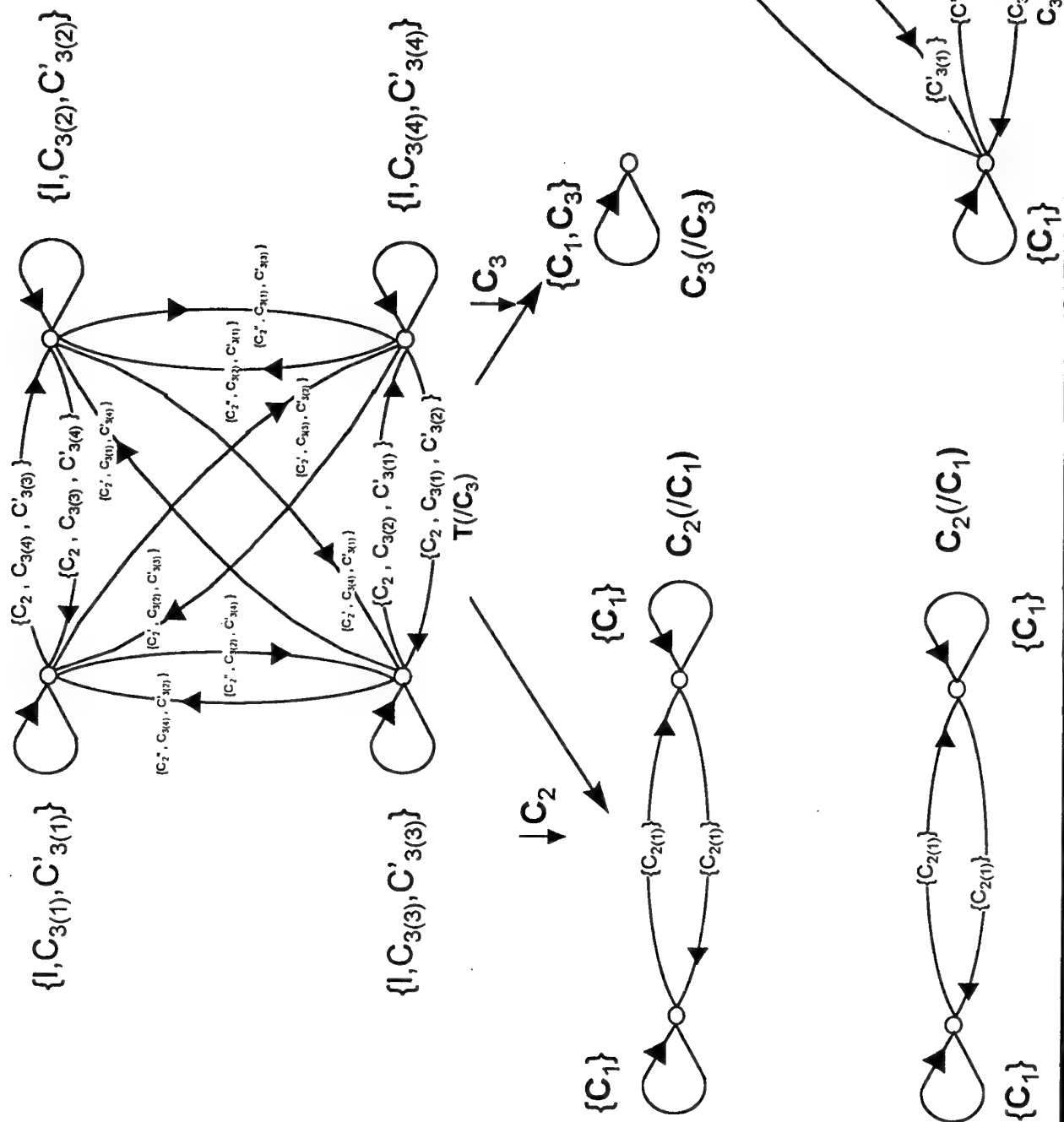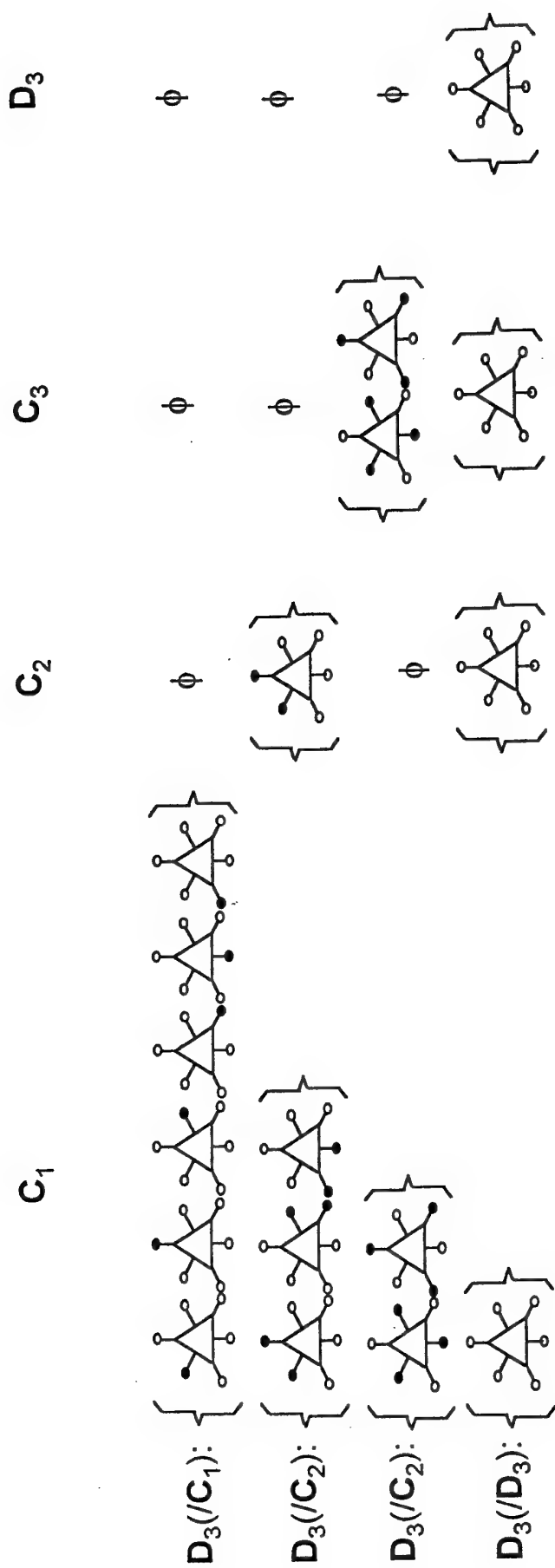
FIG 8
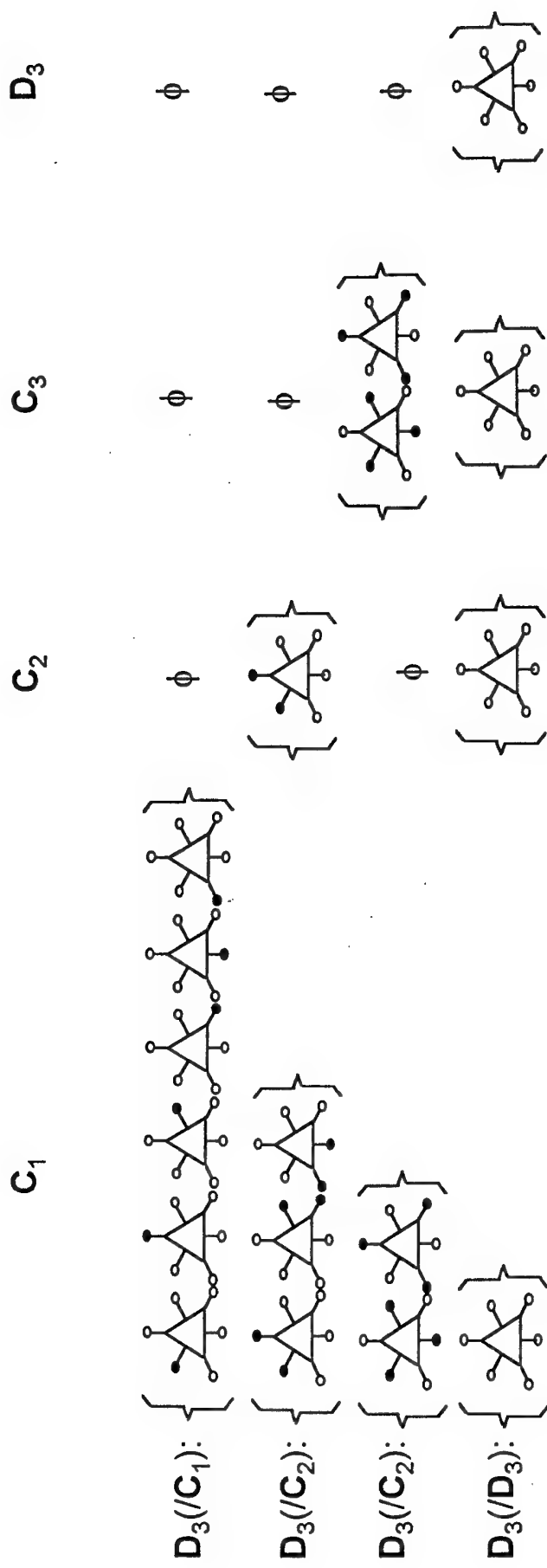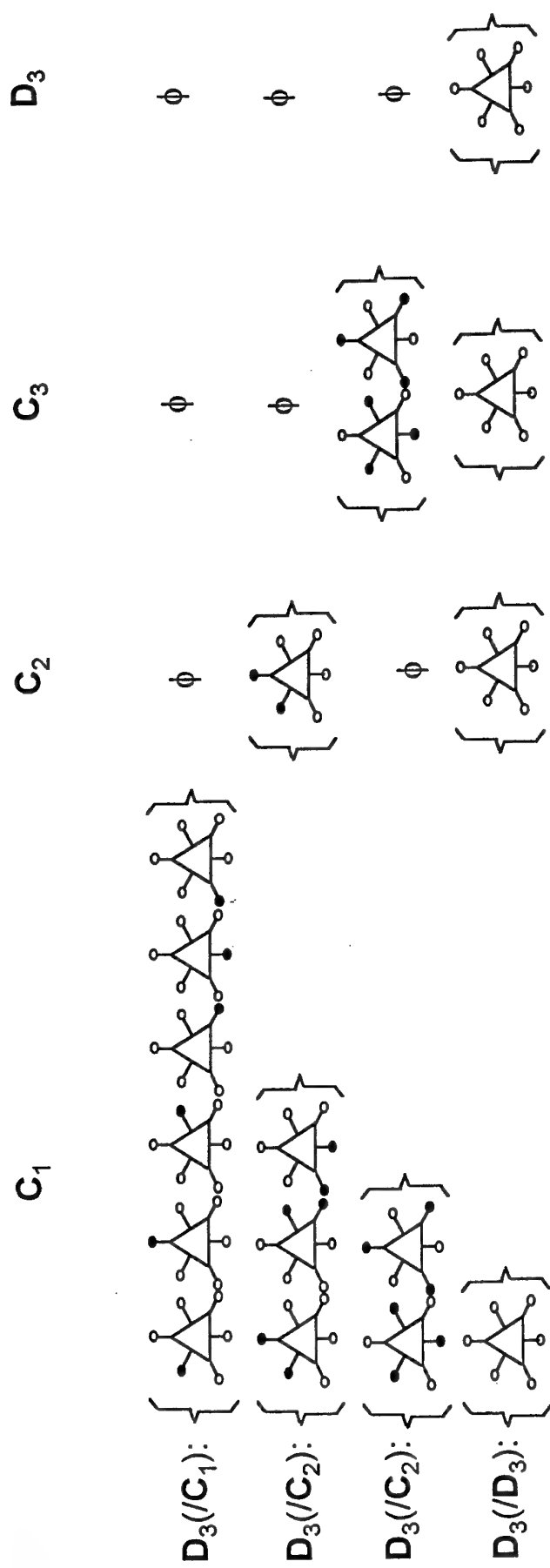
FIG 8

FIG 8
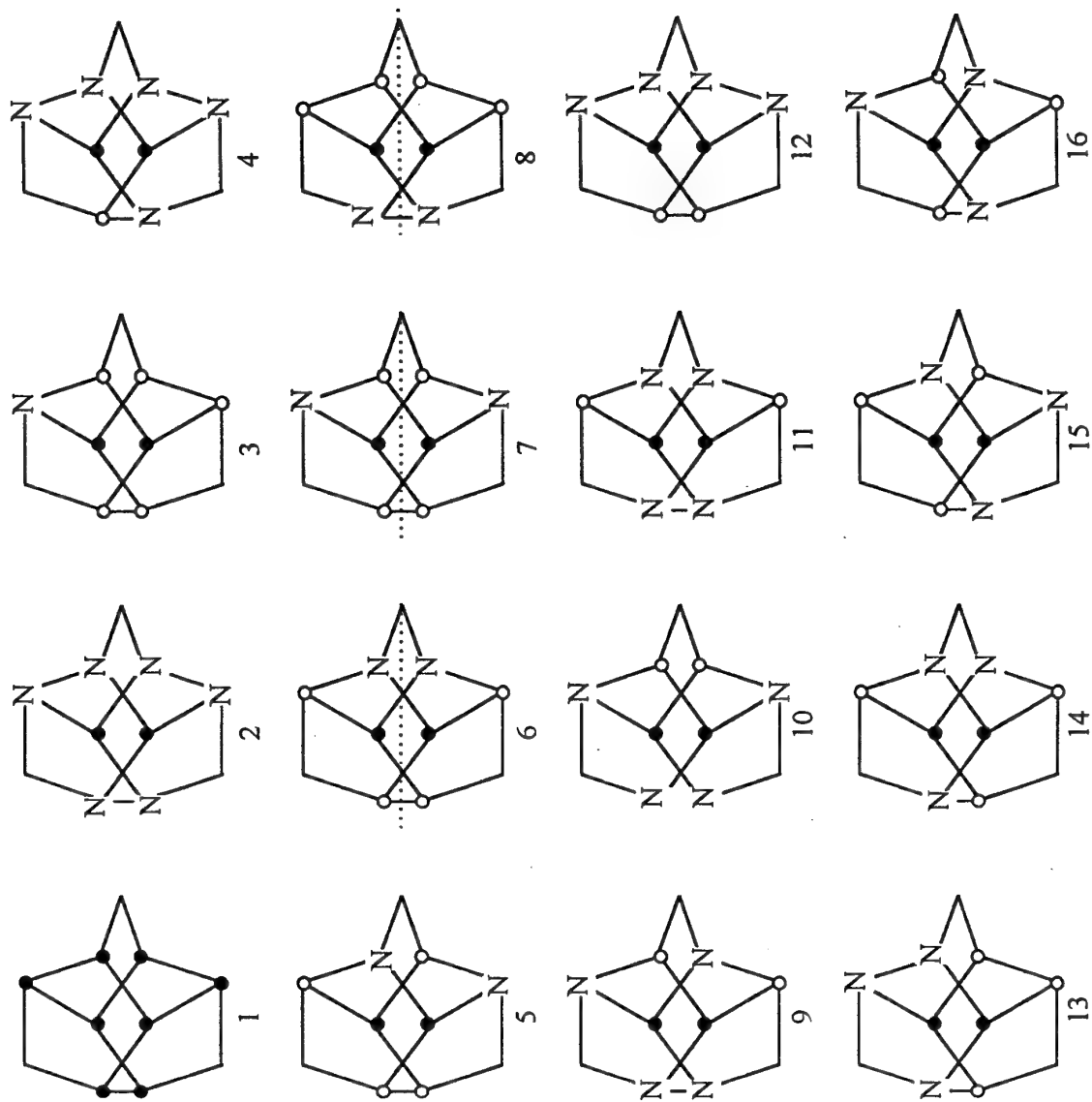
FIG 9

Characterising Graph Drawing with Eigenvectors

Tomaž Pisanski[*]
IMFM, Dept of Theoretical Computer Science,
University of Ljubljana,
Ljubljana, Slovenia
email: `tomaz.pisanski@fmf.uni-lj.si`

John Shawe-Taylor[†]
Dept of Computer Science, Royal Holloway, University of London,
Egham, England
email: `jst@dcs.rhbnc.ac.uk`

July 1999[‡]

Abstract

Two definitions of the problem of graph drawing are considered and an analytical solution is provided for each of them. The solutions obtained make use of the eigenvectors of the Laplacian matrix of a related structure. The procedures give good results for symmetrical graphs and they have already been used for drawing Fullerene molecules in the literature. The analysis characterises precisely what problems the two procedures are solving. It also illuminates why they can perform unsatisfactorily on asymmetrical graphs.

1 Introduction

We consider the problem of embedding a graph on n vertices in Euclidean space $\mathcal{R}^k$, for $k < n$. Typically k would be 3 or 2. By posing the problem

---

1

as minimising the squared norm of the appropriately weighted distance between adjacent points subject to natural normalising conditions we arrive at a formulation of the problem for which the optimal solution can be simply computed in terms of the eigenvectors of the Laplacian matrix of the (weighted) graph. For the case where the weights are chosen to be unity the solution is independent of the uniform penalty given to non-adjacent vertices. In this case and for regular graphs the technique has been applied by Pisanski [9], who demonstrated that the generated drawings are particularly pleasing in the case of Fullerene graphs arising in chemistry. The idea of using eigenvectors for drawing graphs was used first in chemical setting for molecular orbitals; see [8]. A similar technique has been developed by Bolla [2] for generating Euclidean representations of hypergraphs. For distance-regular graphs with a second eigenvalue of multiplicity at least k the embedding has interesting properties; see Godsil [4].

This paper demonstrates that a problem, that has been traditionally solved by gradient descent techniques used to minimise a measure of poverty of the generated embedding, affords an analytical solution which can be implemented in an efficient deterministic algorithm [9]. At the same time it reveals significant insights into the relations between embeddings of graphs and the structure of the eigenspaces of their Laplacian matrices.

The Laplacian matrix has been used in graph embedding before in Tutte's straight line embedding of planar graphs [10, 11]. The approach presented here is related but corresponds to solving the equation without boundary conditions. The characterisation in terms of minimising the sum of distances between vertices is also appropriate in Tutte's case but subject to the chosen cycle being fixed at the boundary, see also Becker and Hotz [1].

Use of eigenvectors to generate embeddings is not new. As early as 1980 Kruskal and Seery [6] devised a sophisticated method for drawing network diagrams using a statistical technique called Multidimensional Scaling (MDS) [5, 7] to arrive at a matrix whose eigenvectors could be viewed as embedding vectors. The approach is closely related to that presented here, but is not characterised in terms of a tightly defined optimization problem.

In Section 4 we discuss in detail the relationship between their method and one of our techniques. It transpires that in certain special cases the solutions obtained by the two methods are up to scaling factors identical. The main advantage of our approach is the theoretical explanation in terms of the two optimization problems which elucidates the strengths and weaknesses of the two methods.

2

2 Notation and Known Results

Let $A(G) = (A_{uv})$ be the adjacency matrix of a simple (positively weighted) n-vertex graph G with no loops. Note that u, v are understood to be adjacent iff $A_{uv} > 0$. For non-adjacent vertices $A_{uv} = 0$. Let D be the $n \times n$ diagonal matrix with non zero entries

$$D_{vv} = \sum_{u:(u,v)\in E(G)} A_{uv},$$

the weighted degree of vertex v. The Laplacian matrix is defined to be $Q(G) = Q(A) = D - A$, where $A = A(G)$.

We summarise a few known results involving the Laplacian matrix. We will number the eigenvalues of $Q(G)$ given in ascending order: $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, with corresponding eigenvectors $j = e^1, e^2, \ldots, e^n$, where j is the all one vector, while $0 < \lambda_2$ if the graph is connected. In addition for any n-dimensional real vector x it can be verified that

$$x^T Q(G) x = \sum_{(u,v)\in E(G)} A_{uv}(x_u - x_v)^2. \tag{1}$$

3 Graph Drawing Problem and Initial Result

We pose the problem of embedding a graph G as finding a mapping

$$\tau : V(G) \longrightarrow \mathcal{R}^k.$$

We will place constraints on this mapping in order to ensure that the representation is natural and hopefully pleasing. We will denote by τ_i the n-dimensional vector formed by taking the i-th coordinate of $\tau(u)$ for all $u \in V(G)$. Thus τ_i is an n-dimensional vector indexed by the vertices of the graph G. Our first requirement is that the centre of gravity of the representation be at the origin. This implies that the vectors τ_i have average entry 0, or $\tau_i \perp j$, for $i = 1, \ldots, k$. The next constraint is that the scaling in all dimensions be similar. This is ensured by requiring that

$$\|\tau_i\|^2 = \sum_{u=1}^n \tau(u)_i^2 = 1.$$

Note that throughout this paper the norm notation $\|.\|$ will as here refer to the 2-norm. Finally we would like the embedding to retain maximum

3

information about the graph. An example of how information can be lost is given when $\tau_i = \tau_j$ for some $i \neq j$, i.e. τ_i and τ_j are maximally correlated. In this case we have effectively reduced the dimension of the representation by one. Hence maximal information will be represented if the vectors have zero correlation, i.e. $\tau_i \perp \tau_j$, for $i \neq j$. We require adjacent vertices to be close together weighted according to A_{uv} (e.g. for different chemical bond types the value might vary), and require non-adjacent vertices to be far apart. Our definition of the graph drawing problem may therefore be stated as follows.

Problem 3.1 *Graph Drawing of an n-vertex graph G given by (weighted) adjacency matrix A in $\mathcal{R}^k$, $k < n$.*

Find a mapping $\tau : V(G) \to \mathcal{R}^k$, which minimises the following energy function

$$E(\tau) = \sum_{(u,v) \in E(G)} A_{uv} \|\tau(u) - \tau(v)\|^2 - \beta \sum_{(u,v) \notin E(G)} \|\tau(u) - \tau(v)\|^2,$$

subject to the constraints

$$\|\tau_i\| = 1, \quad \tau_i \perp j, \quad \text{for } i = 1, \ldots, k$$
$$\tau_i \perp \tau_j, \quad \text{for } 1 \leq i < j \leq k,$$

where β is a positive constant controlling the strength of the force driving non-adjacent vertices apart. ∎

Before proceeding, some further discussion of our problem definition is warranted. Firstly, there seems to be some arbitrariness in the fact that we can specify different 'attractions' between vertices but non-adjacent vertices are all 'repelled' with equal force. We will show that the more general problem created by allowing negative weights can also be solved using the techniques derived for Problem 3.1.

Another aspect of the definition that is a little unsatisfactory is the requirement that the scaling be similar in all directions. Indeed we will see that the method does not work well for highly asymmetrical graphs. In order to avoid this artificial symmetrisation we propose the following second definition of the graph drawing problem albeit with a similar flavour to Problem 3.1.

Problem 3.2 *Graph Drawing of an n-vertex graph G given by (weighted) adjacency matrix A in $\mathcal{R}^k$, $k < n$.*

4

Find a mapping $\tau : V(G) \rightarrow \mathcal{R}^k$, such that the function

$$E(\tau_i) = \sum_{(u,v) \in E(G)} A_{uv} \|\tau(u)_i - \tau(v)_i\|^2 - \beta \sum_{(u,v) \notin E(G)} \|\tau(u)_i - \tau(v)_i\|^2 = 1,$$

for $i = 1, \ldots, k$, while maximising the sum of the norms

$$\sum_{i=1}^{k} \|\tau_i\|^2 = \sum_{u \in V(G)} \|\tau(u)\|^2,$$

subject to the constraints ,

$$\tau_i \perp j, \quad \text{for } i = 1, \ldots, k$$
$$\tau_i \perp \tau_j, \quad \text{for } 1 \le i < j \le k,$$

where β is a positive constant controlling the strength of the force driving non-adjacent vertices apart. ∎

Note that this model allows the τ_i to have different norms, but specifies that a unit length of 'wire' is available in each dimension to create the model. Clearly changing the amount of wire simply has a scaling effect on the solution, so that the problem is well-posed if the number 1 is replaced by any constant. Note also that the requirement also implies that the amount of 'wire' used is the same for all directions since the norms are sums of squares over the coordinates. This observation lends the definition a naturalness that matches the definition of Problem 3.1.

We are now in a position to state our main result.

Theorem 3.1 *Let G be a connected n-vertex weighted graph with adjacency matrix A. The graph drawing problem given in Problem 3.1 is solved by taking the weighted graph with adjacency matrix B with entries*

$$B_{uv} = \begin{cases} (A_{uv} + \beta) & \text{if } (u,v) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

and computing the eigenvectors $e^1, e^2, \ldots, e^n$ with corresponding eigenvalues $0 = \lambda_1 < \lambda_2 \le \ldots \le \lambda_n$ of the Laplacian matrix $Q(B)$. An optimal embedding τ is given by $\tau_i = e^{i+1}$, $i = 1, \ldots, k$ and the minimal value of $E(\tau)$ is

$$\sum_{\ell=2}^{k+1} \lambda_\ell - \beta nk.$$

5

If $\lambda_{k+1} < \lambda_{k+2}$ then the optimal embedding is unique up to orthogonal transformations in $\mathcal{R}^k$.

Corollary 3.1 *In the case where the graph is not weighted (i.e. $A_{uv} \in \{0, 1\}$), the optimal embedding does not depend on the parameter β.*

Proof: If the graph is not weighted and has adjacency matrix A, then $B = (1 + \beta)A$. Hence the Laplacian matrices $Q(A)$ and $Q(B)$ also satisfy $Q(B) = (1+\beta)Q(A)$. This implies that they have the same eigenvectors with the corresponding eigenvalues of $Q(B)$ multiplied by a factor of $1+\beta$. Hence by the theorem the optimal embedding does not depend on the parameter β. ∎

Proof of Theorem 3.1 First note that we can rewrite the energy function $E(\tau)$ as follows.

$$E(\tau) = \sum_{(u,v)\in E(G)} (A_{uv} + \beta)\|\tau(u) - \tau(v)\|^2 - \beta \sum_{(u,v)\in E(K_n)} \|\tau(u) - \tau(v)\|^2, \quad (2)$$

where K_n is the complete graph on the vertices of G with edges weighted 1. If we consider the complete graph in equation (1), the following equality is obtained for an n dimensional real vector x.

$$x^T Q(K_n)x = x^T(nI - J)x = \sum_{u,v\in V(K_n)} (x_u - x_v)^2 \quad (3)$$

where I is the $n \times n$ identity matrix and J is the $n \times n$ all 1's matrix, i.e. $J_{ij} = 1$, for all i, j. In general we have the following relation for an embedding τ and graph G with adjacency matrix A and its Laplacian matrix $Q(A)$.

$$\sum_{(u,v)\in E(G)} A_{uv}\|\tau(u) - \tau(v)\|^2$$

$$= \sum_{(u,v)\in E(G)} A_{uv} \sum_{i=1}^{k} (\tau(u)_i - \tau(v)_i)^2$$

$$= \sum_{i=1}^{k} \sum_{(u,v)\in E(G)} A_{uv}(\tau(u)_i - \tau(v)_i)^2$$

$$= \sum_{i=1}^{k} \tau_i^T Q(A)\tau_i, \quad (4)$$

6

by equation (1). Combining the results of equations (2), (3) and (4), we obtain the following expression for the energy function $E(\tau)$.

$$E(\tau) = \sum_{i=1}^{k} \tau_i^T [Q(B) - \beta(nI - J)]\tau_i \tag{5}$$

Let $j = e^1, \ldots, e^n$ be the eigenvectors of $Q(B)$ with corresponding eigenvalues $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_n$. Without loss of generality we may take $\|e^i\| = 1$ for $i > 1$, since eigenvectors are only determined up to their direction. Note that eigenvectors of a symmetric matrix are orthogonal and so $e^i \perp e^j$ for $i \neq j$. We have

$$[Q(B) - \beta(nI - J)]e^1 = 0,$$

while for $i > 1$, $e^i \perp j$ and so

$$[Q(B) - \beta(nI - J)]e^i = (\lambda_i - \beta n)e^i.$$

Hence the eigenvectors of $Q(B)$ are also eigenvectors of $Q(B) - \beta(nI - J)$. Expressing τ_i in the eigenbasis, we have

$$\tau_i = \sum_{\ell=2}^{n} \mu_i^\ell e^\ell,$$

where $\mu_i^1 = 0$ since $\tau_i \perp j$ $(j = e^1)$. Hence we can write the energy of τ as

$$
\begin{aligned}
E(\tau) \quad &= \sum_{i=1}^{k} \sum_{\ell=2}^{n} (\mu_i^\ell)^2 \lambda_\ell - \beta n k \\
&= \sum_{\ell=2}^{n} \lambda_\ell \sum_{i=1}^{k} (\mu_i^\ell)^2 - \beta n k.
\end{aligned}
$$

The condition $\tau_i \perp \tau_j$ now becomes $\mu_i \perp \mu_j$, while the condition $\|\tau_i\| = 1$ becomes $\|\mu_i\| = 1$. Since the μ_i can be extended to an orthonormal basis matrix M for which M^T is also orthonormal we have

$$\nu_\ell^2 = \sum_{i=1}^{k} (\mu_i^\ell)^2 \leq 1$$

with $\sum_{\ell=1}^{n} \nu_\ell^2 = k$. Hence, the minimum will occur when $\nu_\ell^2 = 1$ for $\ell = 2, \ldots, k+1$ and $\nu_\ell^2 = 0$, for $\ell > k+1$. This can be achieved by taking

7

$\mu_i^{i+1} = 1$, $\mu_i^j = 0$, $j \neq i + 1$, or $\tau_i = e^{i+1}$, $i = 1, \ldots, k$, as stated in the theorem. Note that the minimum energy is

$$\sum_{\ell=2}^{k+1} \lambda_\ell - \beta nk.$$

If $\lambda_{k+2} > \lambda_{k+1}$, then we must have $\nu_\ell^2 = 0$ for $\ell > k+1$ for the minimum to be achieved. This implies that $\mu_1, \ldots, \mu_k$ span the same space as $e^2, \ldots, e^{k+1}$ and can be obtained by an orthogonal transformation of these vectors. Hence the optimal embedding is unique up to orthogonal transformation in $\mathcal{R}^k$. ∎

4   Applications and Further Results

We begin by addressing the problem touched on in the introduction concerning the possibility of solving a problem for which the underlying graph has negative weights.

Theorem 4.1 *Let G be a connected n-vertex weighted graph with some negative weights and adjacency matrix A. The graph drawing problem given in Problem 3.1 is solved by taking the weighted graph with adjacency matrix B with off-diagonal entries*

$$B_{uv} = \begin{cases} (A_{uv} + \alpha + \beta) & \text{if } (u,v) \in E(G) \\ \alpha & \text{otherwise} \end{cases}$$

where $\alpha = -\min\{A_{uv}|(u,v) \in E(G)\} > 0$, and computing the eigenvectors $e^1, e^2, \ldots, e^n$ with corresponding eigenvalues $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_n$ of the Laplacian matrix $Q(B)$. An optimal embedding τ is given by $\tau_i = e^{i+1}$, $i = 1, \ldots, k$ and the minimal value of $E(\tau)$ is

$$\sum_{\ell=2}^{k+1} \lambda_\ell - \beta nk - \alpha k.$$

If $\lambda_{k+1} < \lambda_{k+2}$ then the optimal embedding is unique up to orthogonal transformations in $\mathcal{R}^k$.

Proof: The theorem follows from Theorem 3.1 and the observation that

$$E_{A+\alpha(J-I)}(\tau) = E_A(\tau) + \alpha k,$$

8

which follows from the computations performed in the proof of Theorem 3.1. Hence, a minimum embedding for A is also a minimum embedding for $A + \alpha(J - I)$, while the minimal value of $E(\tau)$ is αk less. ∎

Hence the procedure can also be used to find optimal embeddings of graphs with negative weights as might occur in chemical bonds with different repelling strengths.

A question which might naturally arise when considering a novel embedding strategy is whether it is guaranteed to produce a 2-dimensional drawing with no crossing edges when presented with a planar graph. For the algorithm of Theorem 3.1, this turns out not to be the case as the simple counter-example in Figure 1 shows.
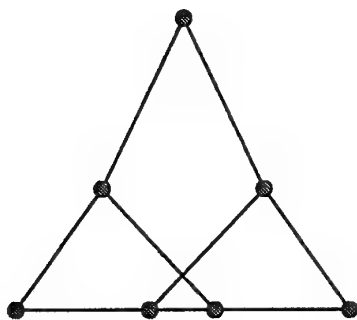
Figure 1: Planar graph drawn with a crossing edge

The graph is C_7 (the cycle on 7 vertices) with two extra edges, $(1,5)$ and $(3,7)$. The graph is clearly planar, but Figure 1 shows the result of applying the algorithm of Theorem 3.1.

A good example of the kind of image generated by our method is given in Figure 2 which is the embedding generated for the Buckminster fullerene in $\mathcal{R}^3$ using the 2nd, 3rd and 4th eigenvectors and taking a two dimensional projection.

9
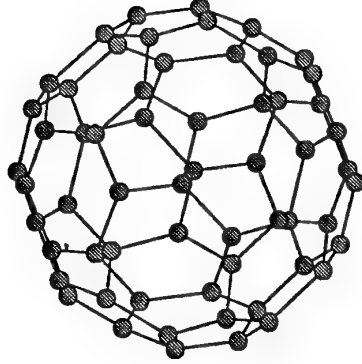
Figure 2: The Buckminster fullerene. The coordinates are determined by the 2nd, 3rd and 4th eigenvector.

In our definition of the graph drawing task (see Problem 3.1), we require that the drawing has normalised variance along the coordinate axes and that the projections onto the coordinates are orthogonal. Together these constraints imply that the drawing will have spherical symmetry in terms of its variance along any axis, since along a (normalised) direction $y = (y_1, \ldots, y_k)$ the variance is

$$
\| \sum_{i=1}^{k} y_i \tau_i \|^2 = \sum_{i=1}^{k} y_i^2 \| \tau_i \|^2
$$
$$
= \sum_{i=1}^{k} y_i^2 = 1
$$

Hence, in a certain sense we are forcing the graph to "look spherical". For graphs with a naturally eccentric shape our method can break down. In order to illustrate this effect, Figure 3 shows how the method draws the Cartesian product of two paths $P_n \times P_m, 2 \le n \le m \le 10$ in $\mathcal{R}^2$. The rows of the figure are indexed by n, while the columns are indexed by $m - n$. Hence the leftmost column contains drawings of $P_n \times P_n$, for $n = 2, \ldots, 10$, while the top row contains the drawings of $P_2 \times P_m$, for $m = 2, \ldots, 10$.
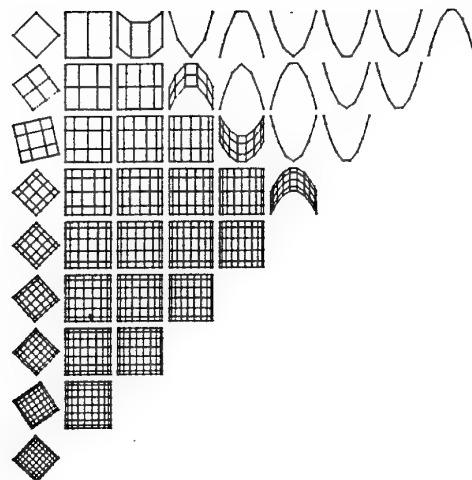
Figure 3: The Cartesian product of two paths $P_n \times P_m, 2 \leq n \leq m \leq 10$, where the coordinates are given by the second and third eigenvector of the Laplacian matrix.

Note that the figures become degenerate when the difference between m and n is too large and both the second and third eigenvalues are inherited from P_m, causing each copy of P_n to map to a point. The method fails to work because the second harmonic in the longer direction corresponds to a lower Laplacian eigenvalue than the first harmonic in the orthogonal direction. If equality of these two eigenvalues occurs then a mixture of the two 'modes' appears in one coordinate, otherwise the second coordinate becomes a quadratic function of the first and the graph drawing collapses onto a line.

In order to show that this problem is not only confined to simple 'two-dimensional' graphs, we include a fullerene graph drawn using our technique, which also possesses a degenerate image (see Figure 4). The graph shown is taken from [8]. Although not immediately apparent from the figure the three-dimensional coordinates of the vertices all lie on a parabolic (two-dimensional) surface, though in this case no pair of vertices is actually given the same coordinates. This explains why in this case a better image is created by taking the 2nd, 4th and 5th eigenvectors, [8, 9], since the third

11

eigenvector is a harmonic of the second.
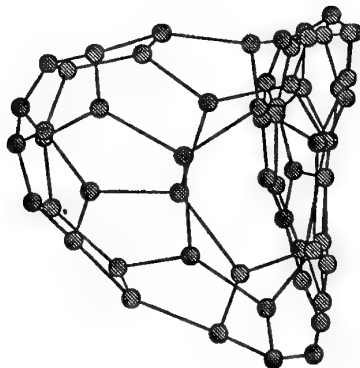
Figure 4: A fullerene on 60 vertices. The coordinates are determined by the 2nd, 3rd and 4th eigenvector.

We conclude this section by presenting a solution to the second Graph Drawing Problem 3.2, which to a certain extent overcomes the enforced symmetry implicit in Problem 3.1.

Theorem 4.2 *Let G be a connected n-vertex weighted graph with adjacency matrix A. The graph drawing problem given in Problem 3.2 is solved by taking the weighted graph with adjacency matrix B with entries*

$$B_{uv} = \begin{cases} (A_{uv} + \beta) & \text{if } (u,v) \in E(G) \\ 0 & \text{otherwise} \end{cases}$$

and computing the eigenvectors $e^1, e^2, \ldots, e^n$ with corresponding eigenvalues $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_n$ of the Laplacian matrix $Q(B)$. An optimal embedding τ is given by

$$\tau_i = \frac{1}{\sqrt{\lambda_{i+1} - \beta n}} e^{i+1}, i = 1, \ldots, k.$$

Proof: Using the analysis of Theorem 3.1 we can write

$$E(\tau_i) = \sum_{\ell=2}^{n} (\lambda_\ell - \beta n)(\mu_i^\ell)^2,$$

12

where

$$\tau_i = \sum_{\ell=2}^{n} \mu_i^{\ell} e^{\ell},$$

and

$$\sum_{i=1}^{k} \|\tau_i\|^2 = \sum_{\ell=2}^{n} \sum_{i=1}^{k} (\mu_i^{\ell})^2.$$

By the observation after the definition of Problem 3.2, the solution will be invariant to orthogonal rotations. We may therefore assume that the τ_i are aligned with the eigenvectors e^{i+1} when projected into the subspace spanned by $e^2, \ldots, e^{k+1}$. Hence, $\mu_i^j = 0$, for $j \neq i+1, j \leq k+1$. It is now clear that

$$\sum_{i=1}^{k} \|\tau_i\|^2,$$

is maximised by also taking $\mu_i^j = 0$, for $j > k+1$, since the norm will have to be smaller in order to have $E(\tau_i) = 1$, after multiplying by larger eigenvalues. Hence, the optimal solution is given by taking $\tau_i = c_i e^{i+1}$. But then

$$\begin{aligned} E(\tau_i) &= (\lambda_{i+1} - \beta n)c_i^2 = 1, \\ \Rightarrow c_i &= \frac{1}{\sqrt{\lambda_{i+1} - \beta n}}, \end{aligned}$$

as required. ■

The algorithm that is proposed in Theorem 4.2 has already been adopted by Manopoulos and Fowler [8] with improved results for less symmetrical graphs than the algorithm of Theorem 3.1.

If we apply the MDS method of Kruskal and Seery [6] to a graph which is vertex transitive (i.e. has a group of automorphisms which acts transitively on the vertices, ensuring the graph 'looks the same' from the viewpoint of any vertex), the result will be that τ_i is set to a different but related multiple of e^{i+1}. This follows from the fact that the MDS method acts as a uniform procedure in this case together with the fact that the eigenvectors of the adjacency matrix of a regular graph coincide with those of its Laplacian matrix.

13

5 Conclusions

We have presented an analysis characterising two graph drawing procedures that have been adopted by different researchers, principally for drawing Fullerene molecules. The characterisation is pleasing in itself, but also throws light on the performance of the procedures and in particular clarifies when they are likely to perform well.

It is not clear how the results might be generalised if the norms used are altered, either in the energy function or in the accompanying constraints on the vectors τ_i. It is likely that an analytical solution will not be possible in this case.

A question that remains unresolved in our understanding of the application of these methods, is a satisfactory way of determining when the eigenvectors for the smallest eigenvalues are harmonics of those already used and should therefore be discarded. Those using the methods have derived various heuristics but it would be useful to gain greater understanding of the factors involved.

References

[1] B. Becker and G. Hotz, On The Optimal Layout of Planar Graphs with Fixed Boundary, *SIAM J. Computing*, **16** (1987) 946–972.

[2] M. Bolla, Spectra, Euclidean representations and clustering of hypergraphs, *Discrete Math.*, 117 (1993) 19–39.

[3] P.W. Fowler, T. Pisanski, and J. Shawe-Taylor, Molecular Graph Eigenvectors for Molecular Coordinates, *In* Graph Drawing, R. Tamassia and E.G. Tollis (eds), Lecture Notes in Computer Science, 894, Springer-Verlag, Berlin, 1995.

[4] C.D. Godsil, Bounding the diameter of distance regular graphs, *Combinatorica*, 8 (1988) 333 - 343.

[5] P.E. Green and V.R. Rao, Applied Multidimensional Scaling: a comparison of Approach and Algorithms. Holt, Rinehart and Winston, New York, 1972.

14

[6] J.B. Kruskal and J.B. Seery, Designing Network Diagrams. Proceedings of First General Conference on Social Graphics, U.S. Department of the Census, Washington, D.C. pp. 22–50, July 1980.

[7] J.B. Kruskal and M. Wish, Multidimensional Scaling. Sage Publications, California, 1978.

[8] D.E. Manolopoulos, P.W. Fowler, Molecular graphs, point groups, and fullerenes, *J. Chem. Phys.*, **96** (1992) 7603–7614.

[9] T. Pisanski, Recognizing Symmetric Graphs, presented at Graph Drawing '93. Sèvres, France, September 26-29, 1993.

[10] W.T. Tutte, Convex representations of graphs, *Proc. London Math. Soc.*, **10** (1960) 304–320.

[11] W.T. Tutte, How to draw a graph, *Proc. London Math. Soc.*, **13** (1963) 743–768.
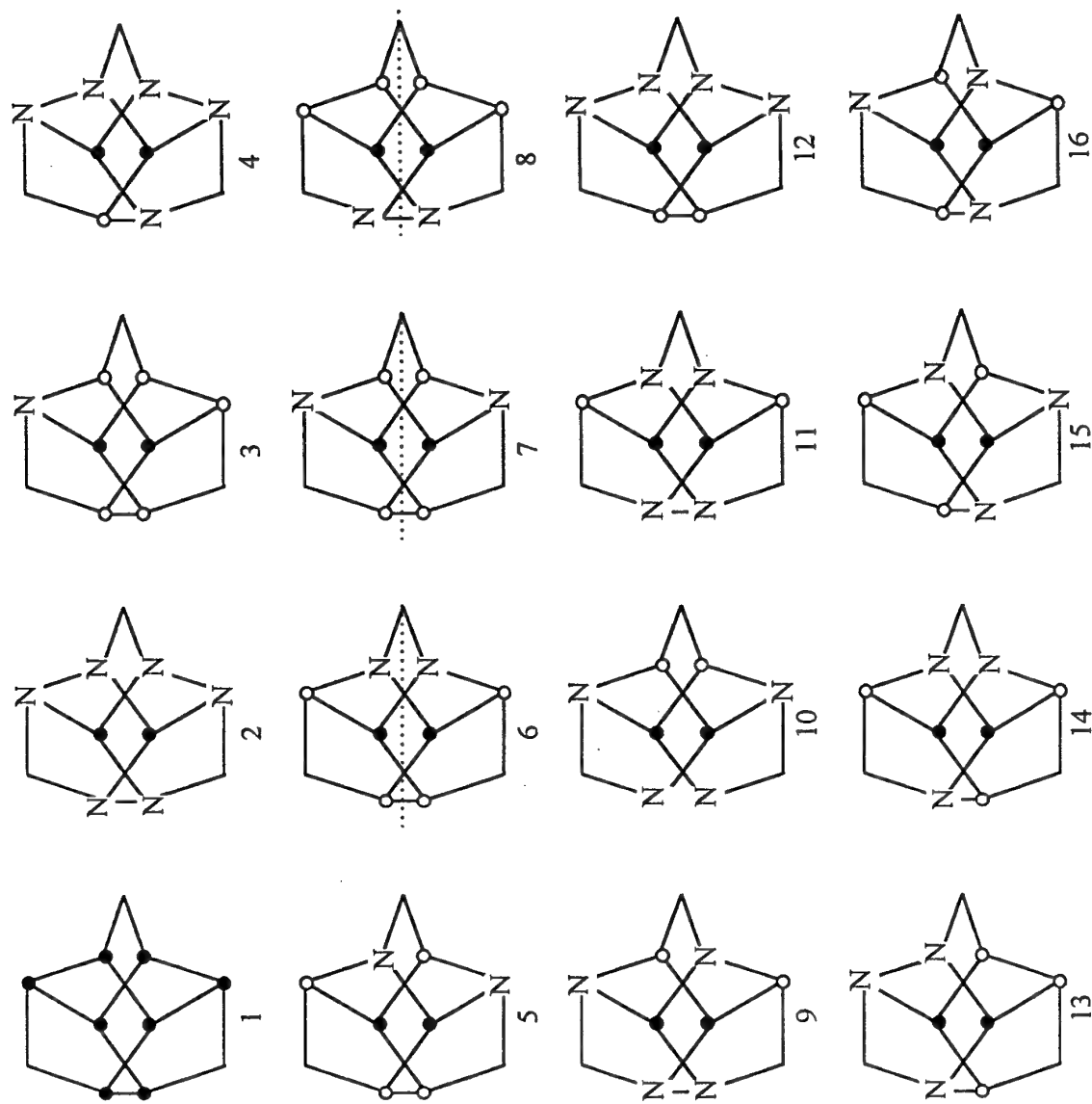
FIG 9

On Numerical Characterization of Cyclicity

Tomaž Pisanski,[a] Dejan Plavšić,[b] and Milan Randić[c]

[a]IMFM, Department of Mathematics, University of Ljubljana,

Jadranska 19, 1000 Ljubljana, The Republic of Slovenia

e-mail: tomaz.pisanski@fmf.uni-lj.si

[b]The Ruđer Bošković Institute, P.O.B. 1016, HR-10001 Zagreb,

The Republic of Croatia

e-mail: dplavsic@rudjer.irb.hr

[c]Department of Mathematics and Computer Science, Drake University,

Des Moines, IA 50311, USA

e-mail: milan.randic@drake.edu

We propose characterizing of the cyclicity of molecular graphs by considering their D/DD matrix. Each non-diagonal element of D/DD is a quotient of the corresponding elements of the distance matrix D and the detour matrix DD of a graph. In particular, we are using the leading eigenvalue of the D/DD matrix as a descriptor of cyclicity and are investigating for monocyclic graphs C_n how this eigenvalue depends on the number of vertices n, as n approaches infinity.

1. INTRODUCTION

The distance matrix D of (molecular) graphs has received considerable attention in mathematical chemistry and has been well studied.[1,2] The elements of the matrix are the distances d_{ij}, where d_{ij} stands for the number of edges on the shortest path between vertices i and j. The sum of the elements above the main diagonal in D gives the Wiener number,[3,4] a well-known graph invariant of interest in structure-property correlations. The *detour matrix DD* of graphs, although suggested in the mathematical literature some time ago,[5] has only recently received attention in mathematical chemistry.[6-10] The elements $(DD)_{ij}$ of the detour matrix are the length of the longest path between the vertices i and j considered. It is interesting to observe that different graphs can have identical detour matrix (when vertices are suitably labelled), the first instance of the situation in graph theory that non-isomorphic graphs are represented by an identical matrix.[11]

1.1 *D/DD* Matrix

Construction of novel matrices and novel graph invariants by using quotients of matrix elements of two different matrices or two different invariants, respectively, has been introduced in chemical graph theory only relatively recently.[12-21] An example is the matrix whose (i, j) elements are obtained as a quotient between Euclidean and topological (graph theoretical) distance between vertices i and j. In this case the leading eigenvalue apparently offers a measure of *folding* (bending) of the long chain.[22]

Recently, a new graph matrix, the *D/DD* matrix, has been introduced.[23] Its diagonal elements are by definition equal to zero and off diagonal elements are given as a quotient of the corresponding elements of the distance matrix D and the detour matrix DD. Note, although the DD matrix can be identical for non-isomorphic graphs, the distance matrix D is different for them and consequently the D/DD matrix has to be different for them as well. It has been suggested that the D/DD matrix might offer a novel characterization of cyclic structures.[23] In particular its leading eigenvalue has

been put forward as an index of *cyclicity* of a graph.

Cyclicity as well as branching are concepts that have been widely used in chemistry in a qualitative fashion. Attempts to assign to such concepts a numerical magnitude have resulted in different definitions for these quantities that are difficult to define. In a way, the prevailing definition is one that has been found useful, either in the characterization of a model or because it leads to a further development of the field. To illustrate the aforesaid let us recall of alternative generalizations of the Wiener number, the well-known graph invariant which was defined initially only for trees, to cycle-containing systems.[13-16] Similarly any definition of cyclicity will be to a degree arbitrary, hence the intention is to come up with a definition that will quantify cyclicity and involve as little as possible non structural arbitrary choices. In order to achieve such a goal we should better understand the cyclicity of simple structures first. Having that in mind we had decided to examine more closely characterization of monocyclic systems. Therefore, we examined how the leading eigenvalue of *D/DD* varies with n, the size of the cycle graphs, C_n, representing simple monocyclic structures.

2. MONOCYCLIC SYSTEMS

Small monocyclic systems C_n, $n = 3 - 8$ are illustrated in Figure 1. Cycle graphs, C_n, are vertex and edge transitive[5] leading to adjacency matrices in which each row can be obtained from the first row by a cyclic permutation. The same applies to the distance and detour matrices of C_n and consequently to the *D/DD* matrix of C_n as well. Hence, cycle graphs have simple *D/DD* matrices. In each row of *D/DD* of an odd cycle C_{2k+1} every element appears twice except the diagonal element (Table 1). In case of even cycles C_{2k+2} in each row of *D/DD* the largest element and the diagonal element appear once whereas the remaining elements appear twice (Table 2). As it is known from linear algebra[24] the leading eigenvalue of a symmetric matrix is bounded from above and from below by its largest and smallest row sum, respectively.

In case of vertex transitive graphs, the two sums are equal, hence for C_n the leading eigenvalue of D/DD equals any row sum. Because the elements of the D/DD matrices of monocyclic systems have such a simple structure it is not difficult to write down the explicit expressions for their row sums, U_n:

$$U_n = \begin{cases} 2 \displaystyle\sum_{k=1}^{\frac{n-1}{2}} \frac{k}{n-k} & \text{if } n \text{ odd} \\[4mm] 1 + 2 \displaystyle\sum_{k=1}^{\frac{n}{2}-1} \frac{k}{n-k} & \text{if } n \text{ even} \end{cases} \tag{1}$$

In Table 3 we have listed the values of the row sums U_n of C_n for smaller values of n. It is easy to see that if n increases to infinity then the row sums U_n also increase indefinitely, even though the increments in each step are getting smaller. In this respect the U_n sums are reminiscent of the harmonic series which is divergent, but at very slow rate. As it is known, the difference between the harmonic series and the logarithmic function of n leads to the Euler constant γ ($\gamma = 0.5772...$). The sequences built from the non zero elements of the first row of D/DD matrices may be also of some interest in mathematics. Consider, for example, first few non zero matrix elements of D/DD for the cycle C_{11} (shown in Table 1) and C_{12} (shown in Table 2):

$$\frac{1}{10}, \frac{2}{9}, \frac{3}{8}, \frac{4}{7}, \frac{5}{6} \quad ; \quad \frac{1}{11}, \frac{2}{10}, \frac{3}{9}, \frac{4}{8}, \frac{5}{7}, \frac{6}{6} \quad .$$

We can observe some resemblance of these sequences with the corresponding harmonic sub-sequence

$$\frac{1}{10}, \frac{1}{9}, \frac{1}{8}, \frac{1}{7}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1 \quad .$$

In the next section we consider the asymptotic behavior of U_n.

3. ON THE CONVERGENCE OF NORMALIZED ROW SUMS

The *cyclicity constant* W_n has been defined in this journal, ref. 23, as the *normalized row sum* $W_n = U_n / n$. The denominator makes the corresponding series convergent. Table 4, and in particular its lower part, indicates that the normalized row sum converges as $n \to \infty$, and that the convergence is slow. For instance, for $n = 20$ two digits, for $n = 100$ three digits, and for $n = 25000$ only eight digits of the limiting value W when $n \to \infty$ are reproduced. Moreover, the normalized row sums W_n for odd n and even n give a lower and an upper bound of W, respectively.

To calculate W with the aforementioned accuracy demands a lot of computations owing to the slow convergence of W_n. Hence, the question arises can the limit be calculated analytically to even higher accuracy. The answer is affirmative since we have found out an expression making this possible. Its derivation shall be stated here.

Let us define

$$S_{m,n} = \frac{1}{n} \sum_{k=1}^{m} \frac{k}{n-k} \tag{2}$$

Using $S_{m,n}$ one can write:

$$W_n = \begin{cases} \dfrac{2}{n} \displaystyle\sum_{k=1}^{\frac{n-1}{2}} \dfrac{k}{n-k} = 2S_{\frac{n-1}{2},n} &, \text{ for odd } n \\[4ex] \dfrac{1}{n} + \dfrac{2}{n} \displaystyle\sum_{k=1}^{\frac{n}{2}-1} \dfrac{k}{n-k} = \dfrac{1}{n} + 2S_{\frac{n}{2}-1,n} &, \text{ for even } n \end{cases} \tag{3}$$

$S_{m,n}$ can be related to harmonic series:

$$S_{m,n} = \frac{1}{n} \sum_{k=1}^{m} \frac{k}{n-k} = \frac{1}{n} \sum_{k=1}^{m} \frac{k-n+n}{n-k} = -\frac{m}{n} + \sum_{k=1}^{m} \frac{1}{n-k} = -\frac{m}{n} + \sum_{t=n-1}^{n-m} \frac{1}{t} \tag{4}$$

· and by shifting the summation index it can be transformed into:

$$S_{m,n} = -\frac{m}{n} + \sum_{t=n-m}^{n-1} \frac{1}{t} = -\frac{m}{n} + \{H(n-1) - H(n-m-1)\} \tag{5}$$

where $H(n)$ is the sum of the initial n terms of the harmonic series:

$$H(n) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \tag{6}$$

$S_{m,n}$ has been represented above as a telescopic sum in which the corresponding members of the two harmonic series that diverge will cancel each other. Formally, we can introduce $\log n$ function in order to convert the divergent harmonic sequences to convergent sequences:

$$S_{m,n} = -\frac{m}{n} + \{H(n-1) - \log(n-1)\} - \{H(n-m-1) + \log(n-m-1)\}$$
$$+ \{\log(n-1) - \log(n-m-1)\} \tag{7}$$

It is well-known that

$$\lim_{n \to \infty} \left[(H(n) - \log n) \right] = \gamma, \tag{8}$$

where $\gamma = 0.5772156649...$ is the Euler constant.

For large value of n the cyclicity constant W_n is approximately equal to $2S_{n/2,n}$, being given by the expression

$$S_{n/2,n} = -\frac{1}{2} + \{H(n-1) - \log(n-1)\} - \{H(n/2-1) + \log(n/2-1)\}$$
$$+ \{\log(n-1) - \log(n/2-1)\}. \tag{9}$$

Finally, we obtain the sought-after limit W in a closed form:

$$W = \lim_{n \to \infty} W_n = 2 \lim_{n \to \infty} S_{n/2,n} = 2(-1/2 + \gamma - \gamma + \{\log(n-1) - \log(n/2-1)\})$$

$$= -1 + 2 \lim_{n \to \infty} \log \frac{n-1}{n/2-1} = 2\log 2 - 1 = 0.38629436112... \tag{10}$$

The advantage of this analytical result for the cyclicity constant of cycle graphs C_n as n tends to infinity is apparent. For instance, in computations for Table 4 we needed 2000 terms to obtain accuracy of about one part per million, because of the slow convergence of the original series. However, if we use the analytical expression it is possible to obtain additional significant digits of W without difficulty. As a mathematical curiosity we note that the Euler constant γ does not appear in the expression for W contrary to what one would expect.

4. ON THE INTERPRETATION OF THE CYCLICITY MEASURE W_n

In Table 4 we listed the difference between W_n for adjacent n values. As n increases the difference is decreasing and tends to zero. In order to arrive at an interpretation of W_n we have to consider what other structural elements of C_n approach zero as n tends to infinity. The curvature, which is in case of circle given by $1/R$, tends to zero as R increases, i.e., as finite segments of circle approach line. W_n is independent of the geometrical scale, thus it cannot have relation to the curvature. We may, however, consider a discrete analogue of curvature defined by the angle θ_n between the sides of a regular n-gon. In contrast to the concept of curvature in geometry which is scale-dependent (see Figure 2) now curvature is scale-independent. Thus, for example, curvature of all hexagons of Figure 2 is constant, while that of concentric circles decreases as R increases.

Discrete curvature is a measure of departure of an n-gon from circle. Clearly, as n increases the difference between n-gon and circle decreases (which has historically been the basis for the early calculation of π). Thus we can take the difference W_n -W (given in Table 5) as a measure of "smoothness" of discretized circles.

An alternative approach is to consider instead of W_n the quantity W_n/n (see Table 5). This quantity has an advantage over W_n , and the difference W_n -W that it does not

alternate with even/odd parity changes, but as we see from Table 5 it monotonically decreases as $n \to \infty$. Clearly W_n/n can be taken as a measure of the "smoothness" of discretized circles.

The above interpretation answers a number of questions that could be raised when considering numerical characterization of cyclic structures. It is clear now why the convergence of W_n is so important. It is not merely a matter of computation, but the approach offers a basis for measuring the "smoothness" of discrete curves. We hope that more light will be brought on characterization of cyclicity by extending the present work to polycyclic systems.

ACKNOWLEDGMENTS

REFERENCES AND NOTES

1. Buckley, F.; Harary, F. *Distance in Graphs;* Addison-Wisley: Reading, MA, 1990.

2. Mihalić, Z.; Veljan, D.; Amić, D.; Nikolić, S.; Plavšić, D.; Trinajstić, N. The Distance Matrix in Chemistry. *J. Math. Chem.* **1992**, 11, 223-258.

3. Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, 69, 17-20.

4. Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Japn.* **1971**, 44, 2332-2339.

5. Harary, F. *Graph Theory*; Addison-Wisley: Reading, MA, 1969.

6. Lukovits, I. The Detour Index. *Croat. Chem. Acta* **1996**, 69, 873-882.

7. Nikolić, S.; Trinajstić, N.; Jurić, A.; Mihalić, Z. The Detour Matrix and the Detour Index of Weighted Graphs. *Croat. Chem. Acta* **1996**, 69, 1577-1591.

8. Lukovits, I.; Razinger, M. On Calculation of the Detour Index. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 283-286.

9. Trinajstić, N.; Nikolić, S.; Lučić, B.; Amić, D.; Mihalić, Z. The Detour Matrix in Chemistry. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 631-638.

10. Trinajstić, N.; Nikolić, S.; Mihalić, Z. On Computing the Molecular Detour Matrix. *Int. J. Quantum Chem.* **1997**, 65, 415-419.

11. Randić, M.; DeAlba, L. M.; Harris, F. E. Graphs with the Same Detour Matrix. *Croat. Chem. Acta* **1998**, 71, 53-68.

12. Randić, M. Search for Optimal Molecular Descriptors. *Croat. Chem. Acta* **1991**, 64, 43-54.

13. Mohar, B.; Babić, D.; Trinajstić, N. A Novel Definition of the Wiener Index for Trees. *J. Chem. Inf. Comput. Sci.* **1993**, 33, 153-154.

14. Randić, M. Novel Molecular Descriptor for Structure-Property Studies. *Chem. Phys. Lett.* **1993**, 211, 478-483.

15. Randić, M.; Guo, X.; Oxley, T.; Krishnapriyan, H. Wiener Matrix: Source of Novel Graph Invariants. *J. Chem. Inf. Comput. Sci.* **1993**, 33, 709-716.

16. Randić, M.; Guo, X.; Oxley, T.; Krishnapriyan, H.; Naylor, L. Wiener Matrix Invariants. *J. Chem. Inf. Comput. Sci.* **1994**, 34, 361-367.

17. Randić, M.; Mihalić, Z.; Nikolić, S.; Trinajstić, N. Graphical Bond Orders: Novel Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1994**, 34, 403-409.

18. D. Plavšić, D.; Šoškić, M.; Landeka, I.; Trinajstić, N. On the Relation between the P'/P Index and the Wiener Number. *J. Chem. Inf. Comput. Sci.* **1996**, 36, 1123-1126.

19. Plavšić, D.; Šoškić, M.; Lerš, N. On the Calculation of the Molecular Descriptor χ'/χ. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 889-892.

20. Plavšić, D. On the Definition and Calculation of the Molecular Descriptor R'/R. *Chem. Phys. Lett.* **1999**, 304, 111-116.

21. Plavšić, D.; Lerš, N.; Sertić-Bionda, K. On the Relation between W'/W Index, Hyper-Wiener Index and Wiener Number. *J. Chem. Inf. Comput. Sci.*, in press.

22. Randić, M.; Kleiner, A. F.; DeAlba, L. M. Distance/Distance Matrix. *J. Chem. Inf. Comput. Sci.* **1994**, 34, 277-286.

23. Randić, M. On Characterization of Cyclic Structures. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 1063-1071.

24. Roberts, F. S. *Applied Combinatorics;* Prentice-Hall: Englewood Cliffs, NJ, 1984.

25. Wolfram, S. *Mathematica, A System of Doing Mathematics by Computer;* 2nd ed.; Addison-Wisley: Reading, MA, 1991.

FIGURE CAPTION

Figure 1. Small monocyclic graphs C_n, $n = 3, 4, 5, 6, 7, 8$.

Figure 2. Continuos and discretized circles of increasing radius.

Figure 1.

Figure 2.

Table 1. The first row of *D/DD* matrices for monocycles C_n, *n* odd. Other rows are obtained by cyclic permutation

n=3	0	1/2	1/2								
n=5	0	1/4	2/3	2/3	1/4						
n=7	0	1/6	2/5	3/4	3/4	2/5	1/6				
n=9	0	1/8	2/7	3/6	4/5	4/5	3/6	2/7	1/8		
n=11	0	1/10	2/9	3/8	4/7	5/6	5/6	4/7	3/8	2/9	1/10

Table 2. The first row of *D/DD* matrices for monocycles C_n, n even. Other rows are obtained by cyclic permutation

n=4	0	1/3	2/2	1/3							
n=6	0	1/5	2/4	3/3	2/4	1/5					
n=8	0	1/7	2/6	3/5	4/4	3/5	2/6	1/7			
n=10	0	1/9	2/8	3/7	4/6	5/5	4/6	3/7	2/8	1/9	
n=12	0	1/11	2/10	3/9	4/8	5/7	6/6	5/7	4/8	3/9	2/10	1/11

Table 3. The row sums U_n for smaller monocycles C_n

n (odd)	U_n	n (even)	U_n
3	1.00000000000	4	1.66666666667
5	1.83333333333	6	2.40000000000
7	2.63333333333	8	3.15238095238
9	3.42142857143	10	3.91269841270
11	4.20396825397	12	4.67705627706

Table 4. The normalized row sums W_n for monocycles C_n of increasing size n, computed by Mathematica[25]

n (odd)	W_n lower bound	n (even)	W_n upper bound	Difference upper bound - lower bound
3	0.33333333333	4	0.41666666667	0.08333333334
5	0.36666666667	6	0.40000000000	0.03333333333
7	0.37619047619	8	0.39404761905	0.01785714286
9	0.38015873016	10	0.39126984127	0.01111111111
11	0.38217893218	12	0.38975468976	0.00757575758
19	0.38491122740	20	0.38754280635	0.00263157895
29	0.38570018343	30	0.38684960872	0.00114942529
39	0.38596573794	40	0.38660676359	0.00064102565
99	0.38624334852	100	0.38634435862	0.00010101010
499	0.38629235310	500	0.38629636112	0.00000400802
999	0.38629386012	1000	0.38629486112	0.00000100100
1999	0.38629423600	2000	0.38629448612	0.00000025012
4999	0.38629434111	5000	0.38629438112	0.00000004001
9999	0.38629435612	10000	0.38629436612	0.00000001000
24999	0.38629436032	25000	0.38629436192	0.00000000160

Table 5. Alternative measures of "smoothness" of descretized n-circles

n	$W_n - W$	W_n/n
3	- 0.05296102779	0.11111111111
4	0.03037230555	0.10416666667
5	- 0.01962769445	0.07333333333
6	0.01370563888	0.06666666667
7	- 0.01010388493	0.05374149660
8	0.00775325838	0.04925595238
9	- 0.00613563096	0.04223985891
10	0.00497548015	0.03912698413
11	- 0.00411542894	0.03474353929
12	0.00346032864	0.03247955748
19	- 0.00138313372	0.02025848565
20	0.00124844523	0.01937714032
29	- 0.00059417769	0.01330000633
30	0.00055524760	0.01289498696
39	- 0.00032862318	0.00989655738
40	0.00031240247	0.00966516909
99	- 0.00005101260	0.00390144796
100	0.00004999750	0.00386344359
499	- 0.00000200802	0.00077413297
500	0.00000200000	0.00077259272
999	- 0.00000050100	0.00038668054
1000	0.00000050000	0.00038629486
1999	- 0.00000012512	0.00019324374
2000	0.00000012500	0.00019314724
4999	- 0.00000002001	0.00007727432
5000	0.00000002000	0.00007725888
9999	- 0.00000000500	0.00003863330
10000	0.00000000500	0.00003862944
24999	- 0.00000000080	0.00001545239
25000	0.00000000080	0.00001545177